CUK- H03539-4-0355072

AMERICAN PHILOSOPHICAL QUARTERLY

④

Edited by

NICHOLAS RESCHER

With the advice and assistance of the Board of Editorial Consultants:

William Alston
Alan R. Anderson
Kurt Baier
Lewis W. Beck
Roderick M. Chisholm
L. Jonathan Cohen
James Collins
James M. Edie
Peter Thomas Geach
Adolf Grünbaum
Carl G. Hempel
John Hospers
Raymond Klibansky
Ernan McMullin, S.J.
Benson Mates
John A. Passmore
Günther Patzig
Richard H. Popkin
Wesley C. Salmon
George A. Schrader
Wilfrid Sellars
Alexander Sesonske
J. J. C. Smart
Manley H. Thompson, Jr.
James F. Thomson
G. H. von Wright
John W. Yolton

VOLUME 4/NUMBER 1 JANUARY 1967

CONTENTS

I. JAAKKO HINTIKKA: *Time, Truth, and Knowledge in Ancient Greek Philosophy* 1

II. ROBERT J. FOGELIN: *Inferential Constructions* 15

III. DAVID BRAYBROOKE: *The Choice Between Utilitarianisms* 28

IV. HERBERT FINGARETTE: *Performatives* 39

V. ARTHUR W. COLLINS: *Philosophical Imagination* 49

VI. ERNEST SOSA: *The Semantics of Imperatives* 57

VII. JON WHEATLEY: *Logical Connection* 65

VIII. REM B. EDWARDS: *Is Choice Determined by the "Strongest Motive"?* 72

IX. WILLIAM J. WAINWRIGHT: *Natural Rights* 79

PUBLISHED BY BASIL BLACKWELL WITH THE COOPERATION OF THE UNIVERSITY OF PITTSBURGH

AMERICAN PHILOSOPHICAL QUARTERLY

POLIC[illegible]

The *American Philosophical Quarterly* welcomes artic[illegible] [illegible]phers of any country on any aspect of philosophy, substantive or historical. However, only s[illegible] [illegible]nt articles will be published, and not news items, book reviews, critical notices, or "discussion [illegible]

Quar / 70 /

MANUSCRIPTS

Contributions may be as short as 2,000 words or as long as 25,000. All manuscripts should be typewritten with wide margins, and at least double spacing between the lines. Footnotes should be used sparingly and should be numbered consecutively. They should also be typed with wide margins and double spacing. The original copy, not a carbon, should be submitted; authors should always retain at least one copy of their articles.

COMMUNICATIONS

Articles for publication, and all other editorial communications and enquiries, should be addressed to: The Editor, *American Philosophical Quarterly*, Department of Philosophy, University of Pittsburgh, Pittsburgh, Pennsylvania 15213.

OFFPRINTS

Authors will receive gratis two copies of the issue containing their contribution. Offprints can be purchased at low cost through arrangements made when checking proof.

SUBSCRIPTIONS

The price *per annum* is six dollars for individual subscribers and ten dollars for institutions. Checks and money orders should be made payable to the *American Philosophical Quarterly*. Back issues are sold at the rate of two dollars to individuals, and three dollars to institutions. All correspondence regarding subscriptions and back orders should be addressed directly to the publisher (Basil Blackwell, Broad Street, Oxford, England).

335078

AMERICAN PHILOSOPHICAL QUARTERLY

Edited by

NICHOLAS RESCHER

With the advice and assistance of the Board of Editorial Consultants:

William Alston
Alan R. Anderson
Kurt Baier
Lewis W. Beck
Nuel D. Belnap, Jr.
Roderick M. Chisholm
L. Jonathan Cohen
James Collins
James M. Edie
Peter Thomas Geach
Adolf Grünbaum
Carl G. Hempel
John Hospers
Raymond Klibansky
Ernan McMullin
Benson Mates
John A. Passmore
Günther Patzig
Richard H. Popkin
Wesley C. Salmon
George A. Schrader
Wilfrid Sellars
Alexander Sesonske
J. J. C. Smart
Manley H. Thompson, Jr.
James F. Thomson
G. H. von Wright
John W. Yolton

VOLUME 4 (*1967*)

PUBLISHED BY BASIL BLACKWELL WITH THE COOPERATION OF THE UNIVERSITY OF PITTSBURGH

AMERICAN PHILOSOPHICAL QUARTERLY

CONTENTS OF VOLUME 4 (1967)

Page

ALDRICH, VIRGIL C.: *On Seeing Bodily Movements as Actions* 222
BECK, CLIVE: *Utterances Which Incorporate a Value Statement* 291
BRAYBROOKE, DAVID: *The Choice Between Utilitarianisms* 28
CASTANEDA, HECTOR-NERI: *Indicators and Quasi-Indicators* 85
COBURN, ROBERT C.: *Persons and Psychological Concepts* 208
COLLINS, ARTHUR W.: *Philosophical Imagination* 49
DANIELS, CHARLES B.: *Colors and Sensations: Or How to Define a Pain Ostensively* 231
DUGGAN, TIMOTHY AND BERNARD GERT: *Voluntary Abilities* 127
EDWARDS, REM B.: *Is Choice Determined by the "Strongest Motive"?* 72
EMMONS, D. C.: *Justice Reassessed* 144
ENGEL, S. MORRIS: *Wittgenstein's "Foundations" and Its Reception* 257
FOGELIN, ROBERT J.: *Inferential Constructions* 15
FINGARETTE, HERBERT: *Performatives* 39
GAUTHIER, DAVID P.: *Moore's Naturalistic Fallacy* 315
GERT, BERNARD, *see* DUGGAN, TIMOTHY AND BERNARD GERT
HANSON, NORWOOD R.: *The Genetic Fallacy Revisited* 101
HINTIKKA, JAAKKO: *Time, Truth, and Knowledge in Ancient Greek Philosophy* 1
HOITENGA, DEWEY J., JR.: *Logic and the Problem of Evil* 114
MCKINNON, ALASTAIR: *"Miracle" and "Paradox"* 308
MARTIN, RICHARD M.: *Facts: What They Are and What They Are Not* 269
MATTHEWS, GARETH B.: *The Inner Man* 166
NAKHNIKIAN, GEORGE: *The Cartesian Circle Revisited* 251
O'CONNOR, D. J.: *Tests for Intentionality* 173
OFSTAD, HARALD: *Recent Work on the Free-Will Problem* 179
PIKE, NELSON: *Hume's Bundle Theory of the Self: A Limited Defense* 159
POLLOCK, JOHN L.: *Mathematical Proof* 238
ROZEBOOM, WILLIAM W.: *Why I Know So Much More Than You Do* 281
SCHLESINGER, GEORGE: *The Probability of the Simple Hypothesis* 152
SMOKLER, HOWARD: *The Equivalence Condition* 300
SOSA, ERNEST: *The Semantics of Imperatives* 57
SWANSON, J. W.: *Religious Discourse and Rational Preference Rankings* 245
WAINWRIGHT, WILLIAM J.: *Natural Rights* 79
WHEATLEY, JON: *Logical Connection* 65
VON WRIGHT, G. H.: *Deontic Logics* 136

American Philosophical Quarterly
Volume 4, Number 1, January 1967

I. TIME, TRUTH, AND KNOWLEDGE IN ANCIENT GREEK PHILOSOPHY

JAAKKO HINTIKKA

I

IN this essay I shall discuss a tacit presupposition, or a group of presuppositions, which seems to lurk behind certain doctrines of Aristotle's and which seems to have been rather widespread in ancient Greece.

A generalization concerning such widespread tacit presuppositions has of necessity something self-defeating about it. If such a generalization is correct, the presuppositions it postulates were shared by the great majority of philosophers and of ordinary people within a culture. If so, there was little occasion for anyone to challenge these presuppositions, to discuss them, or even to bring them out to the open. In such circumstances, not very much direct evidence is likely to be available to show the existence of these presuppositions.

This does not go to show that broad generalizations concerning more or less unconscious ways of thinking within this or that culture are without philosophical interest. In fact, they seem to be worth a great deal more attention than professional philosophers have devoted to them of late. The difficulty I mentioned perhaps explains part of this lack of interest, however. In most cases, it is not very difficult to put forward intriguing suggestions concerning the general features of people's ways of thinking in different cultures or at different periods of intellectual history. The speculative philosophy of history from Hegel onward bristles with such proposals. However, it is usually much more difficult to substantiate them. Often it is rather difficult to connect the more or less implicit *Weltanschauung* which some philosophers of history have thought that they can perceive in the background of the Greek mind with what we actually know about Greek thinkers or of other facets of the Greek civilization.[1] Hence the largely justified qualms of professional philosophers about imputations of implicit general presuppositions to the Greeks.

It seems to me, nevertheless, that a closer study of the most articulate and systematic Greek philosophers may serve a purpose here. Such a philosopher is much more likely to make explicit some of the presuppositions he shares with his countrymen than the majority of these. He may even have to rely on these presuppositions in his philosophical arguments. A careful study of the general presuppositions of an individual Greek philosopher may therefore throw some light on the implicit conceptual presuppositions of the ancient Greeks in general.

We are primarily concerned here with certain general features of Aristotle's philosophical thinking. Some of the assumptions he makes appear to have parallels in other Greek philosophers, and hence occasion the question whether there is something in the common background of all these philosophers to which these assumptions might be related and which might partly explain them. I shall make some suggestions along these lines toward the end of the paper.

II

The group of presuppositions we deal with here is connected with the notion of time.[2] Many presuppositions in this group seem to stem from a

---

[1] These difficulties are illustrated by the scantiness of genuine evidence for those fashionable contrasts between the ancient Greek and the ancient Hebrew *Weltanschauungen* which have recently been effectively criticized by James Barr in *The Semantics of Biblical Language* (Oxford, Clarendon Press, 1961). The theoretical problems involved in inferences from linguistic to cultural data are also discussed, *inter alia*, by Joseph Greenberg in "Concerning Inferences from Linguistic to Nonlinguistic Data," *Language in Culture*, ed. by H. Hoijer (Chicago, University of Chicago Press, 1956); reprinted in *Psycholinguistics*, ed. by S. Saporta (New York; Holt, Rinehart and Winston; 1961) and by Max Black in "Linguistic Relativity," *The Philosophical Review*, vol. 68 (1959), pp. 228–238; reprinted in Max Black, *Models and Metaphors* (Ithaca, Cornell University Press, 1962), pp. 244–257.

[2] The importance of people's attitudes to the category of time as an indication of their value orientation is brought out clearly by F. R. Kluckhohn and F. L. Strodtbeck in *Variations in Value Orientation* (Evanston, Illinois; Row and Peterson; 1961).

CENTRAL LIBRARY

characteristic tendency which permeates many different parts of Aristotelian thought. It is safer to speak of a tendency here than of an assumption, for apparently Aristotle does not consciously choose this way of thought in preference to explicitly formulated alternatives. Rather, he takes this mode of thought as the only natural one, without ever becoming quite clear of the alternatives it might have and certainly without ever articulating the alternatives. If we nevertheless want to formulate this tendency as an explicit assumption we may say that for Aristotle the typical sentences used in expressing human knowledge or opinion are not among those Quine calls *eternal sentences* (or, even among *standing sentences*) but among those Quine calls *occasion sentences*.[3] That is to say, they are not sentences to which we assent or from which we dissent once and for all. They are sentences to which we can subscribe or with which we must disagree on the basis of some feature or features of the occasion on which they are uttered (or written). In particular, the sentences Aristotle is apt to have in mind are *temporally indefinite*; they depend on the time of their utterance.[3a] They may be said to be relative to the moment at which they are propounded. This relation may be implicit, but it may also be made explicit by the occurrence of such "token-reflexive" expressions as "now" or "at the present moment" in the sentence in question. (Among these we have to count also such expressions as "yesterday" or "tomorrow" where another moment or period of time is specified by reference to the present moment.) In a sense they can be fully understood only if we know what moment this "now" is, i.e., when the sentence in question was uttered or is thought of as being uttered. If Aristotle had been asked to give an example of an arbitrary sentence, he might have chosen something like "Socrates is awake" or "Socrates is walking"—and these we may express equally well by "Socrates is now awake" and "Socrates is walking at the present moment."

If Aristotle's tacit assumption is expressed in this way, however, we are already siding with the moderns against him to some extent. What we have said already expresses to some extent the spontaneous reaction of almost all modern logicians and philosophers to the sentences of the kind we are discussing. A modern logician is likely to avoid the use (and the mention) of such sentences as much as possible.[4] They are usually thought of by him as incomplete or indefinite sentences whose "meaning" or "content" depends on the circumstances in which they are uttered or otherwise propounded. Modern philosophers generally prefer not to deal with such temporally indefinite sentences as "Socrates is awake" or "Socrates is walking"; they prefer to discuss and to use instead sentences obtained from these by somehow specifying the time to which they refer independently of the moment of their utterance. The laws of logic are formulated with only or mainly sentences of the latter sort in view, and other procedures are sometimes thought of as being somehow fallacious.

It is not my purpose here to take sides for or against the Aristotelian assumptions as contrasted with the modern ones. It is important, however, to compare them to see how they differ. How would a modern thinker argue for his view that the "content" or "meaning" of a temporally indefinite sentence (say, "It is now raining") varies? One way of doing so might be as follows. Suppose this sentence is uttered on two different occasions, say yesterday and today. Then the facts that make this sentence true or false are different in the two cases. Yesterday it referred to yesterday's weather; today it refers to today's weather. On one day the sentence is verified or falsified independently of its verification or falsification on the other. It may be true on one day and false on the other. All these things are taken to show that the two utterances of the sentence in question cannot carry the same meaning. Although the sentence in the grammatical sense of the word is one and the same, its content or, as it is often said, the *proposition* it expresses on the two occasions is not the same.[5] Hence if we want to have a satisfactory correspondence between our thought and our language, between the logical and the grammatical form, we really must use a different form of words on the two occasions.

---

[3] W. V. O. Quine, *Word and Object* (New York and London; the Technology Press of M.I.T., and John Wiley and Sons; 1960), §9 and §40.

[3a] For a development of the contrast between temporally definite and indefinite statements see N. Rescher, "On the Logic of Chronological Propositions," *Mind*, vol. 75 (1966), pp. 75-96 (see especially pp. 76-78).

[4] Instances are far too numerous for me to give more than a sprinkling of examples here: Bertrand Russell, *An Inquiry into Meaning and Truth* (London, Allen and Unwin, 1940), p. 113; A. J. Ayer, *Philosophical Essays* (London, Macmillan, 1954), pp. 186–187; Nelson Goodman, *The Structure of Appearance* (Cambridge, Mass.; Harvard University Press, 1951), p. 297; Donald C. Williams, "The Sea Fight Tomorrow," in *Structure, Method and Meaning*, ed. by P. Henle (New York, Liberal Arts Press, 1951), pp. 282–306, especially p. 287. Further examples will be given in the course of the discussion.

My list might give the idea that only fairly recent philosophers of an analytic bend of mind favor eternal sentences over occasion sentences. This impression would be incorrect, as Appendix A to A. N. Prior's *Time and Modality* (Oxford, Clarendon Press, 1957) convincingly shows. In fact, in this respect most of our logical tastes go back to the seventeenth century at least. An early version of the doctrine that eternal sentences are superior to occasion sentences is especially interesting in that it was developed in conscious opposition to Aristotle. For Aristotle, the time-relation in a sentence is carried by the verb. This doctrine is closely related to the fact that a typical Aristotelian sentence refers to the moment at which it is uttered or written, for it was the verb that carried the assertoric element in a sentence and was thought of as creating the judgment expressed by the sentence. In contradistinction to this view, it was already held by the authors of the *Port-Royal Logic* that time-reference is not logically speaking a part of the verb. For them, the temporal aspect is, as it were, part of the subject-matter and not created by creating the judgment.

Whether the points I just made are correct or not, Aristotle would not have accepted them. Aristotle would apparently have accepted the doctrine that the sentence "It is raining" is made true or false by different sets of facts accordingly as it is uttered today or yesterday. However, he would not have been worried about the consequence that one and the same sentence may be true at one time and false at another. He would have rejected the notion of a proposition and would have stuck instead to the actual thoughts of the people who uttered the sentence on the two occasions. When doing so, he would have been willing to argue that the thought expressed by the sentence today and yesterday is one and the same. And all this he not only would have been willing to say; he as much as said so quite explicitly:

> For the same statement (*logos*) seems to be both true and false. Suppose, for example, that the statement that somebody is sitting is true; after he has got up this statement will be false. Similarly with beliefs. Suppose you believe truly that somebody is sitting; after he has got up you will believe falsely if you hold the same belief about him. (*Categories* 5, 4a24–29.)[6]
>
> Statements and beliefs, on the other hand, themselves remain completely unchangeable in every way; it is because the *actual thing* changes that the contrary comes to belong to them. For the statement that somebody is sitting remains the same; it is because of a change in the actual thing that it comes to be true at one time and false at another. Similarly with beliefs. (*Ibid.*, 4a34–4b2.)

These quotations show that Aristotle saw no difficulty in combining the two assumptions which to a typical modern thinker are likely to seem incompatible, viz., the assumption that the truth-value of a temporally indefinite sentence changes with time, and the assumption that the sentence in question may nevertheless express one and the same content or proposition or, as Aristotle puts it, one and the same belief or opinion (*doxa*) on the different occasions on which it is uttered or otherwise propounded. Apparently Aristotle did not find anything strange or awkward about his reconciliation of the two assumptions. Later, we shall examine some reasons why he felt this way.

Some comments are in order here. First of all, the authenticity of the *Categories* has sometimes been challenged. Hence the relevance of the passages we just cited is perhaps not beyond doubt. However, even if this work is not by Aristotle himself, it reproduces views current in the Lyceum at a very early date. Hence its testimony can be trusted provided that there are parallel statements in undisputably genuine works of Aristotle's. This is in fact the case. Exactly the same point that was made in the *Categories* recurs in a slightly briefer form in *Metaphysics* IX, 10, 1051b13 ff. Hence it is surely Aristotle's own view. Many of the passages that will be quoted in the sequel constitute evidence to the same effect.

It is also interesting to observe that the view put forward in the *Categories* is not due to Aristotle's desire to rule out recalcitrant facts or to enhance the architectonic neatness of his system. On the contrary, it would have suited the argument of the *Categories* much better if the author could have taken the modern view that the truth-value of opinions and full-fledged sentences never changes. For the point made there is that substances are the only entities that can take contrary attributes at different moments of time and nevertheless remain numerically one and the same (*Cat.* 5, 4a10 ff.). To this doctrine the changing truth-values of opinions and of sentences constituted an unpleasant counter-example. The counter-example was eventually ruled out by means of a rather unsatisfactory manoeuver. (Substances are the only entities that assume contrary attributes because they themselves change; opinions and sentences do this only because the facts they refer to change.) The fact that the modern view was not even considered by the author of the *Categories* although it would have served his purpose perfectly illustrates the hold of the contrary view of him.

---

[5] See, e.g., G. E. Moore, "A Defense of Common Sense," *Philosophical Papers* (London, Allen and Unwin, 1959), p. 35.

[6] The translation is taken from J. L. Ackrill, *Aristotle's "Categoriae" and "De Interpretatione"* (Oxford, Clarendon Press, 1963). In quoting other works of the Aristotelian *Corpus*, I shall normally use the familiar Oxford translation, edited by W. D. Ross, although departures from it will not always be explicitly indicated. In quoting Plato, I shall mostly follow the translations in the Loeb Classical Library.

We may also register the fact that the statements we quoted from the *Categories* are completely categorical. A sentence or a belief remains "completely unchangeable in every way" although the facts it refers to change. Furthermore, nothing is said of beliefs (opinions) and sentences which perhaps never change their truth-values. Hence the author is clearly thinking in terms of temporally indefinite sentences. Because of the parallelisms between the *Categories* and other parts of the Aristotelian *Corpus*, we may infer that the same was the case with the Stagirite.

I conclude, then, that Aristotle saw no obvious difficulties in the assumption that a temporally indefinite sentence expresses one and the same thought or opinion on the different occasions of its utterance. My main suggestion is that he tended to take temporally indefinite sentences as paradigms of all informative sentences. This led him, among other things, to define some of his key notions so as to be applicable only or primarily to temporally indefinite sentences or to "opinions" corresponding to such sentences.

III

The fact that Aristotle preferred temporally indefinite sentences is not belied by the fact that in his syllogistic theory, and in his theory of scientific method which was built on his syllogistic theory, Aristotle frequently says in so many words that a general premiss has to take in all the different individuals of a certain sort, no matter whether they exist now or at some other moment of time. For instance, "All men are mortal" is about all men, present, past, or future.[7] That Aristotle had to stress this feature of the syllogistic premisses may equally well indicate that there was a tendency among his audience and perhaps even in himself to understand these premisses in a different way. That such a tendency really existed is betrayed by the fact that Aristotle frequently uses temporally indefinite sentences as putative examples of syllogistic premisses, contrary to his own explanations. In at least one passage he indicates explicitly that his example is of this sort.[8]

A study of Aristotle's usage also tends to support our view of his notion of truth. He often speaks of what was or will be or would be true to say at some particular time. The following are cases in point:

> For when nothing was separated out, evidently *nothing could be truly asserted* of the substance that then existed (*Met.* I, 8, 989b6–7).
>
> But there are two senses of the expression "the primary when in which something has changed." On the one hand it may mean the primary "when" containing the completion of the process of change (the moment when *it is true to say* "it has changed") . . . (*Phys.* VI, 5, 236a7–9).[9]

Notice also the care with which Aristotle habitually qualifies his statements of the law of contradiction: one and the same sentence cannot be true and false *at the same time*, he says.[10]

Locutions of this sort are by no means infallible evidence. Such locutions could conceivably be used by a philosopher who is not subscribing to the assumptions I have imputed to Aristotle. In connection with his explicit pronouncements on the relation of truth to time, however, they constitute useful circumstantial evidence. Their frequency shows how deeply ingrained the ways of thinking were which we have found in Aristotle.

There are certain passages in Aristotle's works which may seem to call for a different interpretation. Some passages seem to suggest that in addition to temporally indefinite sentences Aristotle sometimes considered sentences containing some kind of more definite time-reference. The evidence is not conclusive, however, but points predominantly in our direction. When Aristotle says that a sentence includes a time-reference, he normally means simply that its verb is in the present, past, or future tense.[11] Such a sentence normally implies a reference to the moment at which it is uttered. There are also passages in which Aristotle has in mind some kind of closer specification of the time to which a sentence refers. But a second look at these passages shows that he is thinking of a specification (of the time of the occurrence of an event) in terms of "the measurable stretch of time from now onwards to that, or . . . from that on to now" (*Phys.* IV, 13, 222a24–28). For instance, in *Parva Naturalia* (*De Memoria* 2) 452b30 ff. Aristotle discusses the difference between "exact" and "inexact" estimate of time in connection with memory.[12] One of his examples is "the day before yesterday," which is a specification of time in relation to the present day. In fact, the context makes it quite clear that what Aristotle is thinking of in this whole passage are exact and inexact specifications of the length of time which separates the remembered event from the present moment (cf., e.g., 452b8 ff.). Hence we again have an instance of a temporally indefinite time-reference.

---

[7] *Analytica Priora* I, 15, 34b17–18; *Analytica Posteriora* I, 4, 73a28–29, and 8, 75b21–36.

[8] At *An. Pr.* I, 10, 30b37–38 Aristotle considers a syllogistic conclusion which is not necessarily true, but is true "so long as" (ἕως) the premisses are true. These premisses thus cannot refer to individuals existing at all the different times, for if they did, their truth-value could not change. An equally revealing example occurs at 31b8–10, where Aristotle uses the terms "sleeping" and "waking" in a counter-example to a syllogism. These are as clear-cut examples as one might wish of terms which apply to an individual *at a time* only. Hence they contradict Aristotle's own admonitions as to how syllogistic premisses ought to be construed.

[9] Similar statements are found, e.g., in *Met.*, IV, 7, 1012a27–28, *Met.*, V, 30, 1025a14–15, *Physics*, VI, 8, 239a28–29.

[10] E.g., *Met.* IV, 3, 1005b19–20, 23–32; 6, 1011b15–18; *Top.* II, 7, 113a22–23; *De Int.* 10, 20a16–18.

[11] See *De Int.* 3, 16b6–18; *De An.* III, 6, 430a30 ff.

IV

The same predilection for temporally indefinite sentences is found in other ancient philosophers, although not always in as explicit a form as in Aristotle. Virtually all the examples of singular sentences which were used by the Stoics as examples and which are preserved to us seem to be temporally indefinite.[13] What is more important, such temporally indefinite sentences are put forward by the Stoics as examples of sentences which are taken to express a complete *lekton.* These complete assertoric *lekta* or in short *axiomata* of the Stoics are in many respects reminiscent of the "propositions" which many modern philosophers postulate as meanings of eternal assertoric sentences.[14] However, *axiomata* differ from propositions in that they are temporally indefinite in the same way as occasion sentences. By saying "writes," one does not yet express a complete *lekton*, we are told by the Stoics, because "we want to know *who* [writes]." Nevertheless, a sentence like "Dion is walking" is said to express a complete *lekton*, in spite of the fact that it leaves room for the analogous question: "*When* is it that Dion is walking?"[15]

From this it followed that the Stoics spoke freely of changes in the truth-value of a sentence and also (more properly) in the truth-value of *lekta.* As is brought out very clearly by William and Martha Kneale, a *lekton* could change its truth-value and even cease to exist.[16] In his list of the different senses of ἀληθής in the Stoics, Benson Mates distinguishes the use of this notion in connection with propositions (Sense I) from its use in connection with sentences that can change their truth-value or, as Mates calls them, propositional functions with a time-variable (Sense II).[17] The distinction does not seem to be motivated, however. In fact, some of Mates's own examples of Sense I are easily seen to involve sentences (or *lekta*) with changing truth-values. Cases in point are found in Diogenes Laertios, *Vitae* VII, 66, where the temporally indefinite sentence "It is day" occurs as an example, as well as in Sextus Empiricus, *Against the Logicians* II, 10–13.[18] The latter passage is not unambiguous by itself; however, its import is brought out when Sextus later returns to the same topic (op. cit. 85, 88–89), using as an example the same temporally indefinite sentence as Diogenes.

One is also reminded here of the famous Stoic controversies concerning the conditions of the validity of implications.[19] This whole controversy is couched in terms of sentences (or *lekta*) with changing truth-values. For instance, Diodorus Cronus held that "if p, then q" is true if and only if q is true whenever p is true.[20] If p and q were temporally definite sentences, this would reduce to our own truth-table definition of material implication. However, the Diodorean doctrine is known to have been directed against Philo's definition which is essentially just our own truth-table definition. Hence Diodorus is clearly presupposing that p and q are temporally indefinite sentences.

---

[12] I am indebted to Professor John W. Lenz for calling my attention to this passage.

[13] Cf., e.g., Benson Mates, *Stoic Logic* (Berkely and Los Angeles, 1961, originally published as vol. 26 of the University of California Publications in Philosophy) which contains a full discussion of the Stoic logic and translations of a number of sources. The closest approximations to genuine exceptions from this type of example that I am aware of are found in Cicero (*De Fato*, ix, 19, and xiii, 30).

[14] For examples of this sort, see, e.g., Mates, *op. cit.*, pp. 96, 113, 118, 121, 123, etc.

[15] See Diogenes Laertios, *Vitae* VII, 63; Mates, *op. cit.*, p. 16.

[16] William and Martha Kneale, *The Development of Logic* (Oxford, Clarendon Press, 1962), pp. 144–146, 153–155.

[17] Mates, *op. cit.*, p. 132.

[18] See the Loeb Classical Library edition of Sextus, vol. 2, pp. 244–247.

[19] See Mates, *op. cit.*, pp. 42–51; Kneale and Kneale, *op. cit.*, pp. 128–138.

[20] See, e.g., Sextus Empiricus, *Against the Mathematicians*, VIII, 112 ff.

Mates explains the Diodorean definition in terms of quantification over a time-variable.[21] This is justified and illuminating, provided that we realize that there is no trace whatever of such a treatment in the Stoics themselves.[22]

V

One of the most interesting things about the assumption we are dealing with is that it helps us to understand some of the most characteristic features of the Greek epistemology. Since these features are found not only in Aristotle but also in many other Greek philosophers, the existence of this connection suggests that we are really dealing with a common tendency of many Greek thinkers.

The most important feature I have in mind is the widespread Greek doctrine that we can have genuine knowledge only of what is eternal or at the very least forever changeless.[23] This doctrine becomes very natural if we consider it as the outcome of two tendencies: (1) A tendency to think of temporally indefinite sentences as typical vehicles of communication; (2) A tendency to think of knowledge in terms of some sort of direct acquaintance with the objects of knowledge, e.g., in terms of seeing or of witnessing them.

I cannot here document the second tendency as fully as it deserves. That there is something here worth being documented is already shown by the facts of the Greek language. One of the common Greek ways to claim that I know was to use the verb οἶδα which literally taken amounts to saying that I *have seen* the thing in question. If we are to believe Bruno Snell, this was not a mere piece of etymology but a fact the speakers of the language were aware of. According to him, "in the Greek language we can frequently discern that the verb εἰδέναι means, to know on the basis of one's own observation."[24] The same applies also to the important word γιγνώσκω, especially in Homer, and to other verbs, too, as pointed out by Snell.[25]

Similar observations have been made often. W. G. Runciman sums up his patient examination of the relevant aspects of Plato's *Theaetetus* as follows: "The general impression left by the *Theaetetus* is that Plato continued to think of knowledge as a sort of mental seeing or touching."[26] This impression is not changed by Runciman's scrutiny of the *Sophist*: "Although Plato says that all statements are either true or false and that all judgments are merely unspoken statements, he does not thereby . . . commit himself to any modification of what we have seen to be his earlier position on the nature and objects of knowledge."[27]

For our purposes it is especially relevant that there was a marked tendency to conceive of the highest forms of knowledge as being somehow analogous to immediate observation as distinguished from mere hearsay. In short, the highest form of knowledge was thought of as being comparable to that of an eyewitness. In the introduction to his edition of Plato's *Meno*, R. S. Bluck says that "the inferiority of ὀρθὴ δόξα to ἐπιστήμη as a state of awareness of the *a priori* is analogous to the inferiority of second-hand information about empirical matters to the certainty of one who has learnt from personal experience."[28] It is also instructive to note that the kind of universal knowledge which Plato ascribes to the soul in his famous doctrine of recollection is explained by Plato (whether metaphorically or not is not at issue here) as being due to earlier personal experience:

> The soul, then, as being immortal, and having been born again many times, and having seen all things that exist, whether in this world or in the world below, has knowledge of them all. (*Meno* 81 c.)

As pointed out by Snell, there is a striking example of this way of thinking in Homer.[29] When he appeals to his omniscient Muses to help him, he does not represent their omniscience as a consequence of superhuman intelligence or of more insight into the laws that govern the events than the humans possess. In Homer's own words,

---

[21] Mates, *op. cit.*, p. 45.

[22] Cf. P. T. Geach's review of Mates in *The Philosophical Review*, vol. 64 (1955), pp. 143–145.

[23] This tendency is one of the most striking characteristics of both Plato and Aristotle, and is found in other ancient philosophers as well. Its role and background in ancient Greek thought does not seem to have been systematically studied, however.

[24] Bruno Snell, "Ausdrücke für den Begriff des Wissens in der vorplatonischen Philosophie," *Philologische Untersuchungen*, vol. 29 (Berlin, 1924), p. 25.

[25] Bruno Snell, *The Discovery of the Mind* (Cambridge, Mass.; Harvard University Press, 1953), p. 13.

[26] W. G. Runciman, *Plato's Later Epistemology* (Cambridge, England; Cambridge University Press, 1962), p. 52.

[27] *Ibid.*, p. 121; cf. p. 125.

[28] R. S. Bluck (ed.), *Plato's Meno* (Cambridge, England; Cambridge University Press, 1961), p. 33.

[29] Snell, *The Discovery of the Mind* (*op. cit.*), ch. 7, especially p. 136.

> Tell me now, Muses that dwell in the palace of Olympus—
> For you are goddesses, you are at hand and know all things,
> But we hear only a rumour and know nothing—
> Who were the captains and lords of the Danaans.[30]

As Snell puts it, "the goddesses are superior to man for the simple reason that they are always at hand, and have seen everything, and know it now. . . ." In fact, their having seen it was for the Greeks almost a prerequisite for their knowing it. Perhaps more significant than this old testimony is the fact that almost exactly the same view is echoed by Plato and Aristotle:

> You both admit, to begin with, that the gods perceive, see, and hear everything, that nothing within the compass of sense or knowledge [therefore?] falls outside their cognizance (*The Laws X*, 901 d; tr. A. E. Taylor; cf. Aristotle, *Poetics* 15, 1454b2–6, and Empedocles fr. 129).

VI

What happens now if this idea of genuine knowledge as an eyewitness's knowledge is applied to the kind of knowledge that can be expressed by means of temporally indefinite sentences? One case seems to be clear: We know the things which are at the present moment within our sphere of perception. But what about things which are not under our present observation? Take, for instance, the piece of putative knowledge expressed by the sentence "There is snow on Mount Olympus" as uttered by somebody who is not within the sight of the famous mountain. On what conditions is a man right who claims to know what this sentence asserts? The Greeks had, if I am right, a tendency to take this question as being equivalent to the question: When does his claim, "I have seen it," amount to a conclusive evidence that the things are now as he says they are? The obvious answer is: Only if the thing in question never changes. Only on this condition does it follow from his earlier observation that things are still as they were at the time of the observation. If the purported object of his knowledge changes, it may have changed in the interval between his seeing it and his making the statement. If the snow sometimes melts from Mount Olympus, then the fact that I have seen the snow there does not go to show that there is snow there now. Only if the snow never melts does the "I have seen it" assertion amount to knowledge concerning the present state of affairs. Hence, the two tendencies I mentioned made it very natural for the Greeks to adopt the view that there can be genuine knowledge only of what is unchangeable (and perhaps also of what is being perceived at the present moment).

This is in fact the way in which Aristotle argues for his doctrine that we have knowledge in the full sense of the word only of what is eternal or forever unchangeable:

> Now what scientific knowledge (ἐπιστήμη) is, if we are to speak exactly and not follow mere similarities, is plain from what follows. We all suppose that what we know is not even capable of being otherwise; of things capable of being otherwise we do not know, when they have passed outside our observation, whether they exist or not. Therefore the object of scientific knowledge is of necessity. Therefore it is eternal. . . . (*Nichomachean Ethics* VI, 3, 1139b20–23.)

It is worthwhile to notice Aristotle's locution "We all suppose." It shows that Aristotle thought that what he was saying was not a peculiarity of his but rather a commonplace among the Greeks.

A concurrent reason why it was easy for Aristotle (and for the other Greeks) to accept the doctrine that there can be knowledge only of what is eternally the same is implicit on his view that sentences like "Socrates sits" express one and the same opinion every time they are asserted. It is natural to say that an opinion of this sort cannot amount to real knowledge if it is sometimes false. For "false knowledge" struck the Greeks, as it strikes us today, as a misnomer.[31] Hence opinions which correspond to temporally indefinite sentences can constitute knowledge only if they are always true, i.e., only if they pertain to facts which never change. And if opinions of this kind are thought of as typical, then one may be inclined to say generally that we can have knowledge only of what is indestructible and unchangeable.[32]

---

[30] Homer, *Iliad*, beginning of the *Catalogue of Ships*.

[31] That it seemed a misnomer to Plato may be gathered from *Gorgias* 454d, from *Republic* 476e, and from *Theatetus* 152c and 186e. Cf. also Parmenides, fr. 2 (Diels), lines 7–8.

[32] This is to all intents and purposes the point of Aristotle's remarks in the *Categories* 7, 7b27–30: "Destruction of the knowable carries knowledge to destruction. . . . For if there is not a knowable there is not knowledge—there will no longer be anything for knowledge to be of. . . ."

If we keep in mind that for Aristotle what is always necessary and that for him what is contingent sometimes will fail to be,[33] we can see that essentially this point is made by Aristotle in *Posterior Analytics* I, 33, 88b31–34:

> . . . that which is necessary cannot be otherwise; but there are propositions which, though true and real, are also capable of being otherwise. Obviously it is not knowledge that is concerned with these. . . .

In many pronouncements of Aristotle's, both these reasons appear intertwined. The following seems to be especially worth quoting:

> For this reason, also, there is neither definition nor demonstration about individual sensible substances, because they have matter whose nature is such that they are capable both of being and of not being; for which reason all the individual instances of them are destructible. If then demonstration is of necessary truths and definition is a scientific process, and if, just as *knowledge cannot be sometimes knowledge and sometimes ignorance*, but the state which varies thus is opinion, so too demonstration and definition cannot vary thus, but it is opinion that deals with that which can be otherwise than it is, clearly there can neither be definition nor demonstration about sensible individuals. For *perishing things are obscure . . . when they have passed from our perception*; and though the formulae remain in the soul unchanged, there will no longer be either definition or demonstration (*Met.* VII, 15, 1039b27–1040a5; my italics).

This may be compared with *An. Post.* I, 6, 74b33 ff., where a similar point is made. Here we obviously have one of the reasons why, for Aristotle, there could not be any genuine knowledge of sensible particulars, but only of universals. In *De Anima* III, 3, 428b8–9 we similarly read: "But true opinion only becomes false when the fact changes unnoticed."

VII

Similar considerations seem to have been operative in Plato, too. In fact, Aristotle attributes exactly the same mode of argument to Plato:

> Plato accepted his [Socrates'] teaching, but held that the problem [of a universal definition] applied not to sensible things but to entities of another kind—for this reason, that the common definition could not be a definition of any sensible thing, as they were always changing (*Met.* I, 6, 987b4–8).

A few lines earlier, Aristotle alleges that Plato accepted the Heraclitean doctrine "that all sensible things are ever in a state of flux and that there is no knowledge about them." Whether these attributions are correct or not, similar juxtapositions of doctrines are found in Plato's writing. From our modern point of view, we may be puzzled and surprised by the facility with which Plato connects in the *Theaetetus* the view that "Knowledge is perception" and the doctrine of Cratylus that all things are constantly changing.[34] This connection is introduced at 152 d, and it is frequently made use of in the argument; witness such passages as the following:

> If all things are forever in motion, every answer to any question whatsoever is equally correct (*Theaetetus* 183 a).

Why should—how could—the constant and universal change postulated by Cratylus make true statements impossible? Clearly only if the statements in question primarily pertain to the moment of their utterance, and to other moments of time only in so far as things remain constant.[35]

The notion of a sentence with a changing truth-value may perhaps serve to explain Plato's strange doctrine of degrees of truth which Runciman finds so puzzling.[36] A sentence cannot at any given moment be truer than another true sentence, but it may be truer than another in the sense of being true *more often* than the latter is. Thus the different degrees of truth are in effect different degrees of unchangeability in the objects these truths are about. This connection is explicitly made by Plato in the *Philebus* (58 a–59 c).

If my interpretation of *De Interpretatione* 9 is correct, Aristotle there in fact uses the expression μᾶλλον ἀληθής meaning "true *more often*."[37]

Another form of the same idea is that an opinion

---

[33] See my paper, "Necessity, Universality and Time in Aristotle," *Ajatus* vol. 20 (1957), pp. 65–90.

[34] For reasons other than the one we are interested in which Plato may have had for connecting the two doctrines, see Norman Gulley, *Plato's Theory of Knowledge* (London, Methuen, 1962), pp. 78–80.

[35] Aristotle, too, assumes that if all things were at rest, "the same things will always be true and false," and that, conversely, if all things were in motion, "nothing will be true and everything will be false" (*Met.* IV, 8, 1012b24–28).

[36] Runciman, *op. cit.*, pp. 124–125.

[37] See Jaakko Hintikka, "The Once and Future Sea Fight: Aristotle's Discussion of Future Contingents in *De Interpretatione* IX," *The Philosophical Review*, vol. 73 (1964), pp. 461–492, especially p. 486.

(*doxa*) is the better the more permanent it is. Thus in the *Meno* (89 c) Socrates says of a certain opinion:

> Yes, but not only a moment ago must it seem correct, but now also and hereafter, if it is to be at all sound.

In the *Republic* (see 430 a) belief or opinion is in the same spirit compared to "a dye, a dye designed to be as 'fast' as possible." The importance of the immutability of the Forms for Plato also becomes intelligible from this point of view. One of the most important roles of the Forms was just to provide absolutely immutable objects of knowledge and thereby secure the possibility of genuine knowledge.[38]

From this point of view we can also appreciate the firm connection which there was in Plato's mind between, on one hand, the distinction between knowledge and true belief and, on the other hand, the distinction between Forms and sensible particulars:

> My own verdict, then, is this. If intelligence and true belief are two different kinds, then these things—Forms that we cannot perceive but only think of—certainly exist in themselves; but if, as some hold, true belief in no way differs from intelligence, then all things we perceive through the bodily senses must be taken as the most certain reality (*Timaeus* 51 d).

VIII

In general, we can see that there is a close connection between two apparently different and even contradictory preoccupations of Plato and Aristotle. On one hand, they paid a great deal of attention to the idea that "Knowledge is perception," whether they in the last analysis accepted it or not. On the other hand, they were attracted by the idea that we can have knowledge only of what never changes. We can now see that these two types of knowledge were both assigned a privileged position by the Greek tendencies to conceive of knowledge as a kind of immediate awareness and to think in terms of temporally indefinite sentences. The two preoccupations are really two sides of one and the same coin.

A reference to our quotation from Homer perhaps helps us to appreciate the appeal of the idea that "Knowledge is perception" to the Greeks. If even the superiority of the divine knowledge was essentially based on the greater share of perceptual evidence that the gods possess, what more could there possibly be to knowledge than perception?

IX

Our observations may also put into an appropriate perspective an argument which Plato's commentators have found puzzling and even mistaken. In *Theaetetus* 201 a-c Plato lets Socrates disprove the suggestion that knowledge can be defined as true opinion by drawing a contrast between the correct opinion of a jury which is giving the right verdict and the knowledge of the event which is possessed by an eyewitness. This may seem puzzling because the objects of which an eyewitness to a crime has knowledge are not the kinds of things of which we can according to Plato's (or Aristotle's) explicit doctrines have genuine knowledge; they are not immutable Forms of which alone we can according to Plato have knowledge in the full sense of the word.

Now the idea that an eyewitness's knowledge is a paradigm case of genuine knowledge was seen to be one of the motives that led to the doctrine that we can have knowledge in the full sense of the word only of what is eternal. This suffices to explain Plato's apparent use of an eyewitness's knowledge as an example of real knowledge without jumping to the conclusion that we could according to Plato have genuine knowledge of sensible reality. The underlying reason why Plato could use the example he mentions is closely connected with the fascination which the idea that "Knowledge is perception" had for him. Of course Plato ended up rejecting this idea, but not before conceiving the highest form of knowledge in analogy to perception, as a kind of "mental seeing or touching," to use Runciman's expression.

One could put the point as follows: What Plato wants to establish in *Theaetetus* 201 a-c is a distinction between knowledge and true belief. For this purpose it suffices for him to show that genuine knowledge is related to true belief in the same way as an eyewitness's "knowledge" is related to "knowledge" by hearsay, the last two being obviously different from one another. Cornford's view that we are here given only an "analogous contrast" has much more to recommend itself than it has recently been given credit for.[39] It is true that Plato does speak of knowledge when he discusses the example, but he seems to restrict the scope of the example in so many words to "matters which one can know only by having seen them and in no other way."

---

[38] Cf. *Parmenides* 135b–c and *Cratylus* 439d–440c.

X

The idea of "changing truth" which we have found in Greek philosophers has to be distinguished from the modern idea of the historical relativity of truth. A historical relativist is apt to argue for an absence of any absolute criteria of truth, which results in an impossibility of obtaining truths that are not liable to be given up when the criteria are changed. What an ancient philosopher like Aristotle had in mind is almost exactly the opposite. He was not concerned with changes in our criteria of truth, but changes in the objects our truths are about. He was not concerned with changes in the opinions we have about reality, but with changes in the reality itself. Aristotle did not think that the discovery of truth is usually very difficult; the difficulty was, rather, that all the truths concerning changing things had to be discovered (as it were) all over again at each new moment.

XI

Our observations concerning the notion of knowledge which naturally results from the Aristotelian presuppositions put these presuppositions themselves into a new perspective. We have so far spoken almost as if a temporally indefinite sentence in the present tense would for Aristotle normally refer merely to the moment at which it is uttered. We can see now that this is not the only way in which Aristotle was willing to understand such sentences and also why he could assume a different interpretation without any compunctions.

The alternative is to understand present-tense sentences in a "tenseless" or perhaps better "conjunctive" sense. In this sense, a sentence will mean that things are *always* in the way the sentence states them to be. A number of passages in the Aristotelian *Corpus* seem to rule out such a sense altogether. *De Interpretatione* 3, 16b6–18 and 10, 19b11–18 as well as *De Anima* III, 6, 430a30 ff. are cases in point. However, in other passages the tenseless sense is clearly presupposed (see for instance, *Topics* V, 3, 131b5–18).

The connection between the notions of time and knowledge which we have pointed out enable us to see that the difference between the two senses was much smaller for Aristotle than we might otherwise expect. For even if a present-tense statement appears to refer exclusively to the moment at which it is made, it can embody real knowledge (as we saw) only if it would have been true to make the same statement at any other moment of time. In so far as present-tense statements (without explicit temporal specifications or explicit restrictions to the present moment) are taken to represent claims to knowledge, they thus automatically assume a "tenseless" or "conjunctive" sense.

The features of Aristotle's thinking which we have noted are related in more than one way to his other logical, semantical, and psychological doctrines. The idea that a temporally indefinite sentence may express one and the same opinion or belief when uttered at different moments of time is encouraged by Aristotle's idea that "spoken sounds are symbols of affections in the soul" (*De Int.* 9, 16a3–4). It is obvious that the sentence, "It is raining," as uttered by me today, is made true or false by a set of facts different from those that verified or falsified my utterance yesterday, "It is raining." But it is very natural to say that in some sense the state of mind or attitude toward my environment that is expressed by the two utterances is the same. The facts which yesterday's utterance refers to are referred to today by the sentence, "It was raining yesterday." But the "state of mind" which this utterance appears to express seems to be entirely different from that expressed by yesterday's present-tense utterance, "It is raining." For instance, the former involves a thought about a temporal difference between the time the sentence refers to and the time of the utterance of the sentence, whereas the latter presupposes no awareness of such a difference.[40] Hence the idea that spoken words are symbols for unspoken thoughts encourages the idea that one and the same temporally indefinite form of words expresses one and the same belief or opinion at the different times when it is uttered.

---

[39] F. M. Cornford, *Plato's Theory of Knowledge* (London, Routledge and Kegan Paul, 1935), pp. 141–142; cf. Runciman, *op. cit.*, pp. 37–38. It may be instructive to observe that in the passage I quoted from Aristotle's *Categories* in n. 32 above he commits an inconsistency somewhat similar to the one Plato seems to me to commit here. Saying that the "destruction of the knowable carries knowledge to destruction" is from an Aristotelian point of view a solecism, for of things destructible we cannot have genuine scientific knowledge (Aristotle's word is here *episteme*) in the first place.

[40] This is borne out by Aristotle's comments in *De Memoria*; see, e.g., 449b24–30. One is here reminded of Russell's paradoxical statement that " 'present' and 'past' are *primarily* psychological terms . . ." (*op. cit.*, p. 113).

XII

It is tempting to express this point by saying that Aristotle presupposed a principle of individuation for propositions different from ours. This way of putting my point turns out to need qualifications, but it is nevertheless useful for many purposes. For instance, we can now see that one of my earlier explanations has to be qualified. The typical sentences considered by Aristotle were said to contain, explicitly or implicitly, a token-reflexive expression like "now" or "at the present moment." This makes the sentences in question in a sense token-reflexive from our modern point of view. The proposition such a sentence expresses is different on different occasions. However, the very same sentences were *not* token-reflexive for Aristotle in this sense, for the belief or opinion they express at different times was for him one and the same.

Notice also that a defender of our modern view cannot argue that Aristotle handles temporally indefinite sentences incorrectly because they express a different meaning or thought (Frege's *Der Gedanke*) in different contexts. Aristotle's procedure implies that we have to individuate thoughts or opinions expressed by temporally indefinite sentences in the same way as these sentences themselves.

XIII

It may be suspected that the peculiarities we have discussed are connected with the general attitudes the Greeks had toward time. It has been suggested that the Greeks "lived in the present moment" to a larger extent than the members of other cultures.[41] It might seem tempting to see in the Aristotelian way of handling the relations of the concepts of time and truth a reflection of the same attitude. It is as if philosophers were so absorbed in the present moment that they tended to think in terms of sentences that contained a reference to the present moment and which therefore dealt primarily with the present state of affairs. It is not easy, however, to find direct evidence for (or against) this suggestion.

There are in any case other general features of the Greek intellectual scene relevant to our subject. Whether or not there was a direct connection between the general Greek attitude toward time and their philosophers' ways of handling this notion, there is the fact that the Greeks were not very successful with their timekeeping—much less successful than some earlier civilizations, not to speak of the Romans. Different cities could have different calendars to such an extent that the year might begin at different times and that the thirteenth month that periodically had to be added to the year was added at different times.[42] Combined with a general neglect of public timekeeping, it is no wonder that the failure of chronology sometimes amounted to a public scandal, as shown by Aristophanes in the *Clouds*. Because of these failures of chronology, there simply was no handy way for the Greeks to take the course modern philosophers generally assume to be the only satisfactory one, viz., to replace all references to the indefinite "now" in a sentence by references to some chronology independent of the moment at which the sentence is uttered. What point would there have been in replacing the sentence "It is raining in Athens today" by a sentence in which the day in question is specified by a reference to a calendar of the form "On such-and-such a day of such-and-such a year it is raining in Athens" if the day and perhaps even the year were different in another city? A reference to the moment of utterance must have been as useful in general as a reference to a badly kept calendar.[43]

---

[41] Cf., e.g., the hyperbolic statements of Oswald Spengler's in *The Decline of the West*, vol. 1 (New York, 1926), p. 131. "Classical man's existence—Euclidean, relationless, point-like—was wholly contained in the instant."

[42] See *A History of Technology*, vol. 3, ed. by C. Singer, E. J. Holmyard, A. R. Hall, and T. I. Williams (Oxford, Clarendon Press, 1954–1958), p. 569.

[43] The connection between the problems of time-measurement and the logic of tenses cuts deeper than modern philosophers sometimes realize. As Y. Bar-Hillel points out in his useful paper, "Indexical Expressions," *Mind*, vol. 63 (1954), pp. 359–379, the problem of converting temporally indefinite sentences into definite ones is not without its presuppositions. To express "The sun is now shining" in temporally definite terms, I must know what time it is (at the place I am talking about). What comes up in the ancient Greek world is primarily the practical difficulty of setting up a useful frame of temporal references. But when we consider statements made at spatially distant points, even difficulties in principle start cropping up. These are of the well-known relativistic variety. If we want to build up a tense-logic which applies to more than one world-line (as is necessary if we, for example, want to study quantified tense-logic) and which is compatible with physical reality, we cannot accept the usual simple tense-logic with a linear structure but must instead have one which has the structure of Lewis' S4. On this point, see my paper "The Modes of Modality," *Acta Philosophica Fennica*, vol. 16 (1963), pp. 65–81, especially p. 76.

It is difficult not to see in this neglect of a systematic calendar a symptom of a general lack of interest in timekeeping and in chronology. It is perhaps not impossible to find some independent evidence of the same attitude. For instance, Hermann Fränkel's study of the early Greek ideas of time has led him to state bluntly that in the *Iliad* "There is virtually no interest in chronology, neither in absolute chronology nor in relative one."[44] He does not find much more interest in later writers, either. This squares rather well with the early Greek ways of expressing temporal relations that Eric Havelock has recently emphasized. He points out that events were located in time not with reference to any absolute chronology but rather by reference to each other: "The basic grammatical expression which would symbolise the link of event to event would be simply the phrase 'and next'."[45]

XIV

It seems to me, however, that there is another closely related feature in Aristotle's background which was more important or at least easier to document. This is the fact that in some obvious though elusive sense the Greek culture was largely based on the *spoken* and not on the *written* word. In philosophical literature this oral character of the Greek thought is shown, for example, by the importance of dialogue as a method of presenting philosophical ideas. One is also reminded of the origins of logic in the technique of oral argumentation. Every reader of Aristotle's *Topics* knows to what extraordinary extent he was concerned with the tricks and pitfalls of verbal exchange.

From this emphasis on the spoken word it follows that the reasons for replacing temporally indefinite sentences by temporally definite ones which modern philosophers have were to some extent absent. If the spoken word is primary in relation to the written word, one is apt to think of and discuss logical and semantical matters from the vantage point of some situation in which the words in question are actually uttered. Now this situation supplies what is missing from a temporally indefinite sentence itself; it enables us to know what the moment of time actually is to which the spoken word "now" refers. A logician eliminates the indefiniteness of temporally indefinite sentences by projecting himself, as it were, to the audience of someone actually uttering the sentence in question. In a written text such words as "now" and "at the present moment" are indeterminate in a sense in which they are not indeterminate as a part of actual speech. In a written culture the replacement of temporally indefinite sentences by definite ones is more important than in an oral culture—to the extent that it might appear to its philosophers as the only "correct" course.

The primacy of the spoken word in relation to the written word is not merely an explanatory hypothesis of a historian; it occurs as an explicit doctrine both in Plato and in Aristotle. Plato's rejection of the written word as a mere aid to one's memory and as being dead and helpless as compared with the spoken word is well-known and need not be elaborated here.[46] Aristotle expresses himself in different terms but equally clearly; according to him "written marks are symbols of spoken sounds" in the same way as the latter were in turn symbols of "affections in the soul."[47]

Eric Havelock has emphasized the same feature in the background of Plato as I am now stressing in the common background of the philosophers of the Socratic school.[48] The use he makes of this idea is different from mine, however. Havelock considers Plato's philosophy as an expression of a transition from a poetic oral tradition to a conceptual culture based on written records. This approach is appealing, but even if the uses Havelock makes of it are fully justified it does not exclude the possibility that many traces of a reliance on the spoken word persisted in Plato's and Aristotle's thinking. If I am right, the heavy reliance on temporally indefinite sentences as a medium of knowledge and of opinion is such a trace.[49]

---

[44] Hermann Fränkel, *Wege und Formen frühgriechischen Denkens*, 2d ed. (Munich, C. H. Beck, 1960), p. 2. Cf. Eric Havelock, *Preface to Plato* (Cambridge, Mass.; Harvard University Press, 1963), pp. 192–193, notes 22 and 27.

[45] Havelock, *ibid.*, p. 180.

[46] Cf., e.g., *Phaedrus* 275d–276a. A full documentation is given in Paul Friedlander, *Platon*, vol. 1 (3d ed., Berlin, Walter de Gruyter & Co., 1964), ch. 5, especially pp. 116–121.

[47] *De Interpretatione* 1, 16a4–6.

[48] Havelock, *op. cit.*, *passim.*

[49] In fact, I would go further and say that in the particular matter at hand, *pace* Havelock, Plato is an arch-conservative who in his logical semantical doctrines relies entirely on the spoken word. The references given in n. 46 above show how deeply Plato felt about this matter.

It is also remarkable that in some unacknowledged sense the spoken word was for Plato and Aristotle even logically prior to the thoughts it expresses. Plato explained the nature of thinking by calling it "the inward dialogue carried on by the mind with itself without spoken sound."[50] This was not merely a metaphor, for Plato felt free to carry out arguments and considerations in terms of spoken sentences and then to transfer the results so as to apply to the corresponding thoughts as well. This is in fact the strategy of Plato's interesting and important discussion at the end of the *Sophist*.

Aristotle does not formulate the logical primacy of the spoken word as compared with thinking as an explicit doctrine. However, he sometimes does exactly the same as Plato, that is to say, he too sometimes bases his views of what we can say of people's *thoughts* on what can be said of their words. The passages quoted above from the *Categories* 5 are instructive cases in point.[51] It may be observed that the word which Aristotle uses in these passages and which was translated by Ackrill as "statement," viz., *logos*, often refers to *spoken* words and sentences.

It is perhaps not irrelevant to mention that both Aeschylus (*Agamemnon* l. 276) and Homer occasionally referred to people's thoughts as "wingless" or unspoken words.

XV

This primacy of concepts applying to spoken words and sentences in relation to concepts referring to thinking must have encouraged the idea that one and the same temporally indefinite sentence expresses one and the same thought (opinion, belief) even when it is uttered on different occasions. If a thought is, logically speaking, nothing but a statement addressed by the "speaker" to himself without spoken sounds, the question whether a statement specifies a complete thought becomes tantamount to the question whether the speaker has to add something to this statement when he tacitly addresses it to himself in order to make it fully understood by the "hearer." The obvious answer is that no expansion is needed as compared with occasions when the same form of word is addressed to someone else, for surely a man understands his own words if they are explicit enough for others to understand them. Hence the grammatical identity of a spoken sentence (form of words) easily becomes a criterion of the identity of the corresponding thoughts, too. Thus a transition from the spoken word to thoughts it expresses does not seem to necessitate any changes in the principle of individuation involved.

The conceptual primacy of the spoken word over thinking is related to the absence of a full-fledged notion of a proposition among the Greeks. By and large, logical and semantical matters were discussed by Greek philosophers in terms of actual utterances rather than in terms of the thoughts or beliefs expressed. It is true that in the opening paragraph of *De Interpretatione* which we discussed the order of relative importance seems to be the opposite: Aristotle says there that spoken sounds are symbols of affections of the soul. However, Aristotle seems to have rather limited purposes in mind in this passage. As Ackrill points out, what the reference to the "affections in the soul" is primarily intended to elucidate is the fact that although spoken and written sounds are different in different languages, what the words express may be "the same for all men," as Aristotle himself says.[52] As Ackrill also points out, "the notion that utterances are symbols of affections in the soul does not have a decisive influence on the rest of *De Interpretatione*." Like other Greek philosophers, Aristotle thought of logical and semantical matters largely in terms of the spoken language.

The same point is applied to Plato by F. M. Cornford:[53]

> Logicians . . . might maintain that there is a false "proposition" . . . which has a meaning, though I cannot believe it. With that we are not concerned, but only with judgments and statements that can be actually made and believed by some rational being. Plato never discusses "propositions" that no one propounds.

Similar remarks can be made of the Stoic concepts of a *lekton* and of an *axioma*, closely though they approximate our notion of a proposition in many other respects. Thus Mrs. Kneale writes as follows:[54]

---

[50] See *Theaetetus* 190a and *Sophist* 263e.

[51] *Categories* 5, 4a24–29, 4a34–b2.

[52] Ackrill, *op. cit.*, pp. 113–114.

[53] Cornford, *op. cit.*, p. 113.

[54] W. Kneale and M. Kneale, *op. cit.*, p. 156.

> As a previously quoted passage has shown, the Stoics wished to insist that an *axioma* which is to be described as true or false must somehow be present when it is so described; ... in one place Sextus clearly assumes that *lekta* exist only when they are expressed or meant.

As far as Aristotle is concerned, it would not be strictly true to say he never considers propositions that no one propounds. He comes very close to envisaging one in *De. Int.* 9, 18b36. It is no accident, however, that this passage is, if I am right, exactly the passage in which Aristotle came closest to considering statements which refer to a unique singular event and which therefore need an objective chronology.[55] The other similar cases that there are in Aristotle of unasserted propositions are much less clear, and probably explainable as mere *façons de parler*.

In the Aristotelian *Physics* we find a neat metaphysical projection of the semantical idea we have discussed. According to the latter, the belief expressed by a sentence containing the word "now" remains ordinarily one and the same. In the *Physics*, Aristotle argues that the "now" is in the actual sense of the word *always the same*. It is what "holds the time together" and makes it continuous. (See *Physics* IV, 13, 222a10 ff., and 11, 219b10–220a4.) There seems to be an interesting analogy between the way different moments of time are actualized by becoming in turn identified with the eternally identical "now," and the way in which the content of a known "now"-statement becomes relevant to the world at the different moments of time by becoming utterable at the moment in question.

It is thus literally the case that the word "now" was not token-reflexive for Aristotle, for each actually uttered "now" referred to the same forever identical actual "now" that is postulated in the *Physics*.

XVI

If there is anything to the contention that the Greeks "lived in the present" to a larger extent than the members of some other tradition, the justification (and perhaps also the import) of this suggestion has to be spelled out in terms of such concrete facts as those that are known about the ways of Greek philosophers with the concepts of time and truth. Whether or not the facts we have pointed out go very far toward establishing any broad generalizations, in any case they show us how the different facets of the Greek way of thinking are related to each other. The suggestion that the Greeks were immersed in the present moment more deeply than we are easily provokes a reply which points to the preoccupation of the Greek philosophers with the eternal and the immutable. We have seen, however, that this preoccupation is closely connected with the reliance of these philosophers on temporally indefinite sentences. Far from being a counter-example to the temporality of Greek thought, this pursuit of the eternal is more of a manifestation of this very temporality—or perhaps rather an attempt to compensate for it. Plato's and Aristotle's ideal of knowledge was knowledge of eternal truths just because the vehicles by means of which truths were thought of by them as being expressed tended to make all other kinds of truths ephemeral. In a sense I am thus led to agreement with the view of R. G. Collingwood that "the Greek pursuit of the eternal was as eager as it was, precisely because the Greeks themselves had an unusually vivid sense of the temporal."[56]

University of Helsinki
and
Stanford University

Received November 4, 1965

---

[55] Hintikka, "The Once and Future Sea Fight," especially pp. 465–466.

[56] R. G. Collingwood, *The Idea of History* (Oxford, Clarendon Press, 1946), p. 22.

American Philosophical Quarterly
Volume 4, Number 1, January 1967

II. INFERENTIAL CONSTRUCTIONS

ROBERT J. FOGELIN

LINGUISTS have remarked that subjunctive constructions are in the process of dying out of our language, and if this is true, it lends urgency to the enterprise of giving a proper account of their character. In this essay I shall give—or at least sketch—such an account by developing the more general notion of an *inferential construction.* My starting point will be Ryle's classic " 'If,' 'So,' and 'Because' "[1]; sorting out the right and the wrong of this essay will get me under way.

I

Ryle presents his main theses in the following paragraph:

> In using "if-then" sentences and in using "because" sentences we are stating or asserting. "If today is Monday, tomorrow is Tuesday" is a true statement; "Tomorrow is Tuesday, because today is Monday" is another statement which may be true. But "Today is Monday, so tomorrow is Tuesday" is not a statement. It is an argument, of which we can ask whether it is valid or fallacious; it is not an assertion or doctrine or announcement which we can ask whether it is true or false.[2]

This passage contains three independent claims, each demanding its own justification:

1. "if-then" sentences are used for making assertions.
2. "because" sentences are used for making assertions.
3. "so" sentences are used to formulate arguments and arguments are not assertions.

Of the three theses only the third is novel, for it alone runs counter to the general presumption (no longer quite so general) that *any* sentence in the indicative has an assertive function. As it is, I think Ryle's radical third thesis is correct but his conservative first thesis is mistaken. The second thesis is, I think, correct but it will take a good deal of preparatory work before we can examine it directly.

I shall not reproduce Ryle's elaborate argument in behalf of the third thesis, for it seems to rest upon two points that are entirely safe:

(A) Arguments are not assertions.
(B) Constructions containing the term "so" as their major connective formulate arguments.

The first claim has behind it the long tradition of subtracting points from student papers where arguments are called true or false rather than valid or invalid. I see no reason to challenge this tradition. In defense of the second claim, we may simply notice that when people present arguments in the ordinary way, they use such connectives as "so," "thus," "therefore," etc. To say much beyond this is to invoke reasons that are less obvious than the conclusion they are intended to establish.[3]

Under the assumption that Ryle's third thesis is correct, we can turn our attention to his first thesis. Why does Ryle maintain that hypothetical sentences express assertions? His argument comes to this:

> Suitability for what may be summarily called the "premissory job" is one of the main things that make us reserve the title "statement" for some sentences in distinction from all the rest.[4]

And:

> Hypothetical statements . . . [unlike arguments] . . . can be premisses and conclusions in arguments (though only in relatively high-level arguments).[5]

No argument is offered in behalf of this claim, which suggests that Ryle finds his position so patently correct as to need no further justification. After all, hypothetical sentences appear as premisses in arguments involving *modus ponens* and as both premisses and conclusions in hypothetical syllogisms. In the face of this strong presumption to the contrary, I shall argue that it is far from clear that "if-then" constructions do perform a premissory job, and in order to build a case for this seemingly unlikely claim I shall examine some of Ryle's own remarks about the relationship between "if-then" sentences and arguments.

---

[1] Gilbert Ryle, " 'If', 'So', and 'Because'," *Philosophical Analysis,* ed. Max Black, (Ithaca, Cornell University Press, 1950).

[2] *Ibid.,* p. 324.

[3] I think that this is true of Ryle's appeals to grammatical considerations which contain bothersome snags. His strongest argument is that any attempt to treat arguments as assertions will lead to the Lewis Carroll paradox of Achilles and the Tortoise; I recommend this line of reasoning to anyone who is uneasy about the force of the brief argument presented above.

[4] Ryle, *op. cit.,* p. 326.

[5] *Ibid.*

On Ryle's own principles, "Today is Monday, so tomorrow is Tuesday" cannot entail "If today is Monday, then tomorrow is Tuesday," for arguments, not being assertions, cannot entail anything whatsoever. What then is the relationship between them? Ryle answers this question in the following passages:

> The argument "Today is Monday, so tomorrow is Tuesday" is an application of "If today is Monday, tomorrow is Tuesday," and it is in this notion of application that lies the answer to our question "How does a valid argument require the truth of the corresponding hypothetical statement?"[6]
>
> . . . the inference "*p, so q*" does not embody "*If p, then q*" as a component of its premiss, but rather applies it in being an operation with "*p*" and "*q*" executed in conformity with it, . . .[7]
>
> . . . no premisses are used or conclusions drawn by a person making a hypothetical statement. But also, much as the operations to be done with ingredients are specified by the very similar but empty handed operation of the demonstrator, *so the inferential operations to be done with statements are specified by the very similar but empty handed operation of the author of a hypothetical statement.*[8]
>
> . . . the author of a hypothetical statement is neither using nor mentioning any premiss statements or conclusion statements. He is showing, empty handed, how to use them. We might say that *he is making a dummy inference*, with a didactic purpose.[9]

Whether "if-then" sentences are *ever* used as premisses in an argument remains to be seen; in any case, this is *not* the role Ryle assigns to them when they are *applied* in an argument. In fact, Ryle's remarks suggest a wholly new way of looking at hypothetical sentences: instead of drawing a contrast between arguments and assertions (and thereby forcing ourselves to include "if-then" sentences in the latter category), we can introduce a new category that calls attention to their points of contact. I shall call this category "inferential constructions." Arguments (e.g., "*p*, so *q*") and inference tickets (e.g., "If *p*, then *q*") are but two species of inferential constructions. The basic feature of an inferential construction is that it contains a dialectical move from one proposition to another. In an argument we make a dialectical move in order to affirm one proposition *on the basis* of another that we also affirm. With an "if-then" sentence we express only the dialectical move and thereby supply a pattern for a potential argument. *It is in this sense that an "if-then" sentence is applied in an argument.*

Nowhere in Ryle's discussion of the relationship between "if-then" sentences and arguments do we find a case of the former playing the role of a premiss in the latter. All that remains is to ask if they *ever* play this role. If we turn to textbooks on logic, we do find entries of the following kind:

$$\begin{array}{l} \text{(a) If } p\text{, then } q. \\ \quad p \\ \hline \therefore q \end{array} \qquad\qquad \begin{array}{l} \text{(b) } p \text{ implies } q. \\ \quad p \\ \hline \therefore q \end{array}$$

We can notice straight off that (b) is ill-formed. The sentence frame ". . . implies . . . " does not accept sentences, but instead, that-clauses or sentence names. It would be better to rewrite it correctly using that-clauses throughout:

$$\begin{array}{ll} \text{(b}') & \text{that } p \text{ implies that } q. \\ & \text{that } p \text{ is true.} \\ \hline & \therefore \text{that } q \text{ is true.} \end{array}$$

What are we to make of the relationship between implication sentences of the form "that *p* implies that *q*" and hypothetical sentences of the form "if *p*, then *q*"? An almost casual remark near the end of Ryle's essay brings out the difference most clearly:

> I have said nothing about statements of the kind " '*p*' *entails* '*q*' " or " '*q*' *does* (or *does not*) *follow from* '*p*' ." Such locutions are used (roughly) not by players in the field but by spectators, critics, and selectors in the grandstand. They belong to the talk of logicians, cross examiners, and reviewers.[10]

---

[6] *Ibid.*, p. 329.

[7] *Ibid.*, p. 331.

[8] *Ibid.*, p. 338 (my italics).

[9] *Ibid.*, p. 339 (my italics).

[10] *Ibid.*, p. 340.

In this list we should also include sentences of the form "that p implies that q"[11] for this will bring out the difference between such constructions and those of the form "if p, then q." From what has gone before, the difference between these two constructions should be obvious: with an "if-then" sentence we perform a dialectical operation; with an "implies" sentence we comment upon the propriety of such an operation. I suggest, though I shall not here defend this suggestion, that when logicians employ entries of type (a), they disregard the distinction between "if-then" sentences and "implies" sentences as it exists in the language. Perhaps the faulty symbolism of (b) provides the halfway house for this transition from "implies" sentences to "if-then" sentences. There is, however, no need for linguistic dogmatism on this matter. A conciliatory move would be to grant the possibility that "if-then" sentences are amphibolic: at times they are used as assertions equivalent in meaning to "implies" sentences; at other times they are used non-assertively as inference tickets. Nothing whatsoever will turn upon the question of whether both of these uses exist, for in what follows I shall limit my attention to the inference ticket use of "if-then" sentences.

Something more should be said about the contrast between assertions and inferential constructions. If I may use the phrase, this is a matter of depth grammar, not surface grammar. We can imagine a language, perhaps it's our own, where inference tickets are called true when the dialectical move they express is legitimate. Similarly, we can imagine a language, which is certainly not our own, where arguments are called true if they are sound. If this were so, then inference tickets and arguments would have one of the standard marks of assertions, but this should hardly disturb us since the term "true" is highly promiscuous and as often as not attaches itself to things that are not assertions. In any case, this single similarity would not overshadow the shared features of arguments and inference tickets that have led to the introduction of the notion of an inferential construction.

II

It has taken some time to arrive at a few modest conclusions, but in philosophy, as elsewhere, we should be governed by the maxim: little steps for little people. The next task is to delineate the differences between arguments and inference tickets. At first blush this seems easy enough; the person who argues "p, so q" thereby commits himself to the truth of both 'p' and 'q', but he is not so committed when he presents the inference ticket "if p, then q." One way to describe this situation is to say that in "p, so q," 'p' and 'q' play an assertive role, while in "if p, then q" they play a non-assertive role.[12] But this will not do, for *nothing* is asserted by the construction "p, so q."

Another way of describing the situation avoids just this difficulty: in the argument "p, so q," one *affirms* "q" on the basis of the proposition 'p' that one also *affirms*. Thus 'p' and 'q' perform a double role in an argumentative construction; they are objects related in a dialectical relationship and they are objects of affirmation. How is affirming related to asserting? If I assert something I thereby affirm it, but there are a variety of ways in which I can affirm something without asserting it. If someone asks me if I am busy, and I reply with the single word "yes," I have affirmed that I am busy, but I have not asserted this. Similarly, I can deny something without asserting its denial—a job commonly done by the word "no." In a parallel fashion the word "so" introduces an affirmational dimension to the dialectical move between 'p' and 'p' and 'q' and it is in this way that we become committed to the truth of both 'p' and 'q' while employing a non-assertive construction. To reflect both the dialectical and affirmational elements in an argument, we can employ the following notational device:

$$p^{+}\rightarrow q^{+} \text{ FOR } p, \text{ so } q$$

The exponents indicate that the two propositions are affirmed and the arrow indicates that a dialectical move has been made from the first to the second. The expression on the left is thus not a statement about an inference, but is a representation of an inference.

We can get a clue about the function of "if-then" sentences from a passage in G. E. Moore's *Commonplace Book:*

> By saying "If p, then q" you generally *imply* that you neither know that p nor that ∼p, nor that q nor that ∼q, . . . Can we say that if it turned out that you did know that both p and q, then what you said, when you

---

[11] I use this form in preference to Ryle's quotation-laden "'p' implies 'q'."

[12] This suggestion is made by William and Martha Kneale in *The Development of Logic* (Oxford, The Clarendon Press, 1962), pp. 50 ff.

> said "if p, then q" was *false*? I think not, though we might say you were not *justified* in saying it, because you would mislead people.[13]

Moore seems to be saying the following: in using an "if-then" sentence we are neither affirming nor denying the component propositions; instead, we are assigning a problematic status to them. The standard reason, but not the only reason, for such an ascription is that we simply do not know the truth status of the component propositions; thus the use of this construction when we do know the truth status of the component propositions would normally be misleading. This suggests the following extension of the notational convention, now using o to indicate a problematic ascription:

$$p^0 \rightarrow q^0 \text{ FOR if } p, \text{ then } q.$$

Guided by these conventions we can now enter the dark waters of subjunctive conditionals. Consider the following specimen sentence:

> If Stevenson had become President, he would have attended cabinet meetings.

We shall assume that this is a counterfactual conditional, i.e., one that commits us to the falsity of the antecedent. How does this commitment come about? As a first attempt we might try the following symbolization:

$$(\sim p^+ \rightarrow \sim q^+)$$

But this will not do, for no inferential move is made from the claim that Stevenson did not become President, even though the sentence as a whole commits its user to just that fact. Instead we can extend the symbolism in the following way:

$$(p^- \rightarrow q)$$

Thus in a counterfactual conditional we make a dialectical move from a proposition that we simultaneously deny, but we do not assert the denial of that proposition, and, *a fortiori*, we do not make a dialectical move from its denial. (The lack of an exponent for the consequent in this inferential construction marks an area for further investigation.)

The systematic point of introducing distinctions between asserting a proposition and affirming it and between asserting the denial of a proposition and denying it comes into focus with a contrary-to-fact conditional. Consider how we would evaluate the claim:

> If Stevenson had become President, he would have attended cabinet meetings.

We accept this claim because being President guarantees attendance at cabinet meetings. Thus the construction is grounded in a covering universal of the following kind:

> If X is President, then X attends cabinet meetings.

But here the 'X' in question is Stevenson, and that amounts to saying that the logical antecedent of this construction is the proposition (expressed in the indicative): Stevenson became President. The locution "Stevenson had become President" is not the label for some curious quasi-proposition; instead, it provides a way for denying a proposition that plays the role of the antecedent in an inferential construction.

The initial decision to distinguish asserting from affirming was suggested by Ryle's remarks about arguments. We now see that this approach admits of a natural extension that encompasses subjunctive conditionals, and while this hardly certifies the approach, it certainly encourages us to pursue it further.

III

For want of better terminology I shall say that the exponents '+', '—', and 'o' express epistemic commitments. The next natural question is this: Are there other inferential constructions that express epistemic commitments? To answer this we can construct a table showing every possible combination of these values and then we can scout about for dicta in the natural language that correspond to the various combinations:

	Antecedent	Consequent	Inferential Construction
I	+	+	Arguments, e.g., Since p, q
	+	o	Does not arise
	+	—	Does not arise
II	o	+	Indicative-subjunctive conditional
III	o	o	Indicative conditional
	o	—	Does not arise
IV	—	+	Subjunctive conditional
V	—	o	Subjunctive conditional
VI	—	—	Subjunctive conditional

Before I proceed, a few restrictions and a warning are in order. For now I shall exclude modal terms from both the antecedent and the consequent since such terms can interact with the expressed epistemic values in complex ways. I shall also exclude sentences where either the antecedent or the consequent has a complex mood construction, as in "If what you have in your hand were a penny, then it would be made of copper." Needless to say, such constructions where part of the antecedent is in the indicative and part in the subjunctive demand special consideration. The warning is this: the table merely provides a guide for turning up constructions which introduce a dialectical relationship between propositions and at the same time express epistemic commitments. The table can not be read from right to left, for the constructions on the right may perform a variety of jobs in our language. I am making only the modest claim that these grammatical constructions, whatever else they might do, can formulate inferential constructions that ascribe epistemic values to their component propositions.

---

[13] G. E. Moore, *Commonplace Book*, ed. Casimir Lewy (London, Allen & Unwin, 1962), pp. 391–392.

Some areas for further investigation lie in the three rows which have not been graced with a Roman numeral and these, I claim, are not reflected in constructions found in the English language. This is not surprising, for such constructions would have a decidedly peculiar character. A construction corresponding to row two would have the following form:

$$(p^{+}\rightarrow q^{0})$$

When a person makes a dialectical move from a proposition that he affirms to some other proposition, he thereby presents grounds for affirming the second proposition as well. Now it might be possible for a person to make such an inferential move without at the same time recognizing that he has presented grounds for affirming the consequent, but we should hardly expect to find a construction whose special purpose in the language is to facilitate this lack of comprehension. A parallel argument can be given concerning a construction corresponding to row three and row six.[14]

We are now left with the six rows that have been assigned Roman numerals. A brief look at each of these constructions will provide an opportunity for some informal remarks that will set the stage for further developments.

I. $(p^{+}\rightarrow q^{+})$. This kind of construction is used for the formulation of arguments—we affirm one thing on the basis of something else that we affirm. We offer arguments for a variety of reasons: sometimes to *justify* the conclusion, at other times to *offer a basis for explaining* what is asserted in the conclusion. The difference between a justification and an explanation will be the subject of further inquiry, but for the moment we can get along with an intuitive understanding derived from examples. I can *justify* the claim that the generator is working by pointing out that the light is on:

> Since the light is on, the generator is working.

In contrast, I can *explain* why the light is on by pointing out that I flipped the switch:

> The light is on because I flipped the switch.

We have yet to come to grips with "because" sentences. Eventually I shall maintain that in presenting an explanation we are not presenting an argument, although the truth of an explanation will often depend upon the soundness of a counterpart argument. The rationale for this way of speaking will emerge later, but for now we can simply note that arguments are involved in both the process of justification and the process of explanation.

II. $(p^{0}\rightarrow q^{+})$. To the best of my knowledge, this sort of subjunctive conditional has never been fully understood. An example will exhibit its distinctive character. You are standing before a fire that is burning with a vivid green flame and by way of suggesting an explanation for this phenomenon you remark:

> If someone put copper sulphate in the fire, then it would be burning with a green flame.

Here the antecedent is in the indicative while the consequent is in the subjunctive and it is through just this combination of moods that we are able to express the given set of epistemic commitments.

III. $(p^{0}\rightarrow q^{0})$. I have maintained that in using an indicative conditional as an inference ticket, we ascribe a problematic status to both the antecedent and the consequent. We sometimes do this, as Moore suggests, because we are in doubt concerning the status of the antecedent and consequent; at other times we use this construction for rhetorical reasons. We can further notice that a construction of this kind must ascribe a problematic status to the consequent, for if an affirmation of the consequent is appropriate, we should use a construction of the second kind, and if a denial of the consequent is appropriate, we are faced with a construction that does not arise.

---

[14] David Pears has offered a table similar to the one given above, but he was able to fill in eight rows instead of six. For row two he offers "Since this is a raven, it must be black." He then adds, "Here 'must', though strong, is bluff, and leaves the truth value of the consequent open." The word "must" surely is bluff, so bluff in fact that it cancels out the force of the dialectical move. This is the sort of thing we say when faced with the breakdown of a standard piece of reasoning. For row six he offers "If this had been black, its being black would have been a reason." This is in fact an epistemic remark and an instance of row nine. David Pears, "Hypothetical," *Analysis*, vol. 10 (1950), pp. 62–63.

IV–VI. On the table we find that the last three rows correspond to subjunctive conditionals. The common feature of these three rows is that they ascribe the value '—' to the antecedent. When subjunctive conditionals are used to make this ascription, we can call them *counterfactual* conditionals. As we shall see later, subjunctive conditionals have other functions that will interest us, but until further notice we shall concentrate upon their use in the formulation of counterfactual inferential constructions.

With a counterfactual conditional the antecedent is denied, but the consequent can conceivably have any of the three epistemic values. Thus in the English language the epistemic status of the consequent is not reflected in a distinctive grammatical construction. This might suggest that the epistemic status of the consequent is of no importance and for this reason it is submerged in the grammar of our language. Unfortunately, this is too facile an account of a complicated matter. In many cases the force of a subjunctive conditional will depend directly upon our understanding of the truth status of the consequent. Why this is so will become clear in the next section.

IV

The next task is to examine the *functions* of these six inferential constructions and here a few technical notions will prove helpful. I shall call those constructions that involve an *o* in their symbolization *open constructions*, and those that do not, *closed constructions*. I shall also speak of some constructions as being *less open* than others; constructions of types II and V are less open than constructions of type III, and *a fortiori*, any closed construction is less open than any open construction.

Using this terminology we can lay down a principle that will guide us in examining these constructions:

> One function of an open inferential construction is to offer a pattern for a less open construction.

The idea of an open inferential construction *offering a pattern* is obviously related to Ryle's notion of "if-then" sentences playing the role of an inference ticket. In fact, the task of this section is to elucidate the way in which inference tickets function.

Let me illustrate the principle just laid down. A proposition whose truth status is problematic may, at any time, lose this problematic status, and when this occurs a shift takes place in the inferential construction in which it can be appropriately used. I shall call such a shift from an open inferential construction to a less open inferential construction the *cashing in* of an open inferential construction. We can begin with a construction of type III:

(1) If the light goes on, then the generator is working.

Now suppose that the light does indeed go on, then a person noting this is in a position to cash in his inference ticket for an argument. One way of expressing this argument is by means of the following sentence:

Since the light is on, the generator is working.

Given the initial principle concerning the function of an open construction, we can now proceed in a perfectly abstract manner and trace out all possible cashing in relationships between these constructions. This is presented in the following diagram:

$$\begin{tikzcd}
 & (p^{+} \rightarrow q^{+}) & & (p^{-} \rightarrow q^{-}) & \\
(p^{0} \rightarrow q^{+}) \arrow[ur, "p"] \arrow[drr, "\sim p"'] & & (p^{0} \rightarrow q^{0}) \arrow[ul, "p"'] \arrow[ur, "\sim q"] \arrow[ll, "q"] \arrow[rr, "\sim p"'] & & (p^{-} \rightarrow q^{0}) \arrow[ul, "\sim q"'] \arrow[dll, "q"] \\
 & & (p^{-} \rightarrow q^{+}) & &
\end{tikzcd}$$

Here the feathered arrows indicate the direction of cashing in; they do not represent implication or entailment. The letter adjoining each arrow shows the change that allows the cashing in. The arrows, we might say, represent the direction of inquiry. Inference tickets are offered in anticipation of their being cashed in for arguments; arguments are not offered in anticipation of their being devalued to inference tickets. Abstractly, changes in both directions are possible, but when we are concerned with the function of these constructions, it is the movement in the direction of closure that interests us.

At first glance it might seem that cashing in an open construction for a subjunctive construction is a net loss, for why should we be interested in an inferential construction when we think that its antecedent will never be realized? There are at least two reasons why such constructions are interesting and important. The first, which is usually overlooked, is that a dialectical move from a supposedly false proposition to a second supposedly false proposition is sometimes used for presenting a *refutory argument.* The second is that we sometimes use these constructions for formulating *speculations.*

To illustrate the refutory use of subjunctive conditionals let me alter a previous example. Once more we are standing before a fire, but this time it is decidedly *not* burning with a green flame. In this setting we might make the following remark:

> If someone had put copper sulphate on the fire, then it would be burning with a green flame.

Through the use of this construction we have formulated a refutory argument against the claim that someone did put copper sulphate on the fire. Alternatively, we could have used a construction of the first type in order to formulate a justificatory argument:

> Since the fire is not burning with a green flame, no one threw copper sulphate on it.

Thus there is a close kinship between refutory arguments formulated in constructions of type VI and justificatory arguments formulated in constructions of type I. They are, in fact, related *via* transposition in a fashion that will later be developed in some detail.

Only constructions of the fourth kind fall completely outside the justificatory-refutory network, for with the epistemic values $(p^- \to q^+)$, they can never be used to formulate arguments that justify the consequent nor be used to formulate arguments that refute the antecedent. We can expect, then, that they will only be used in speculative contexts where we are not interested in justifying or refuting anything. We shall return to the speculative use of subjunctive conditionals after an extended excursion into the domains of covering laws and explanations.

V

Evidently there is some close relationship between generalizations and the six implicational constructions that we have been examining. Given the generalization "All crows are black," any of the following inferential constructions is warranted provided that the epistemic commitments are in order:

I. Since h is a crow, h is black.
II. If h is a crow, then h would be black.
III. If h is a crow, then h is black.
IV.–VI. If h were a crow, then h would be black.

There seem to be two ways in which we might describe this relationship between generalizations and singular inferential constructions. We can identify the generalization "All crows are black" with an inferential construction of the following sort:

If x is a crow, then x is black.

Here generality is captured by using a variable instead of a constant within the inferential construction. This seems to be what Ryle is getting at when he describes generalizations as "season inference tickets." On the other hand, we might refuse to *identify* generalizations with season inference tickets and instead say that generalizations provide the foundation for an inferential construction containing variables. For our present purposes, it does not seem to matter how we decide this issue for it is the move from a general inferential construction to a singular inferential construction that will concern us. For my part, whenever I speak of a covering law, I shall be thinking of a general inferential construction and I shall invariably use this kind of construction in my discussion.

In the literature we often find a distinction drawn between two kinds of generalization: nomological generalizations and generalizations of accidental correlation. On the basis of this distinction it is then argued that some definitive relationship exists between nomological generalizations and singular subjunctive conditionals. I wish to clear away one roadblock to the understanding of subjunctive conditionals by showing that this simply is not so.

We can begin with a generalization that we shall assume asserts only an accidental correlation:

Everyone on the bench is Irish.

This provides the basis for a universal inferential construction of the following kind:

If x is on the bench, then x is Irish.

We know that Jean-Paul Sartre is not Irish, and

335078

this, I suggest, warrants the following instantiation into a singular subjunctive conditional:

> If Jean-Paul Sartre were on the bench, then he would be Irish.

One way to certify this instantiation is to read the subjunctive conditional as a *counter-identical*:

> If Jean-Paul Sartre were identical with someone on the bench, then he would be Irish.

But I wish to show that this instantiation holds even if we do not give the subjunctive conditional the peculiar reading of a counter-identical.

The first step is to give this construction an appropriate setting to work in. Suppose I have independent evidence that everyone on the bench is Irish—I have asked them and each has replied *yes* in an authentic Irish brogue. Someone then tells me that Jean-Paul Sartre is on the bench and I reply:

> If Jean-Paul Sartre were on the bench, he would be Irish.

I have thus formulated a refutory argument. Alternatively, we can transpose the universal inferential construction and then instantiate it into the first kind of singular inferential construction:

> Since Jean-Paul Sartre is not Irish, he is not on the bench.

These two ways of formulating the argument come to exactly the same thing: any justificatory argument in the indicative can be converted into a refutory argument in the subjunctive, and the converse relationship holds as well. In sum, a subjunctive conditional expressing a refutory argument can be derived quite nicely from a generalization expressing only an accidental correlation.

Why, then, do people resist the inference from "Everyone on the bench is Irish" to "If Jean-Paul Sartre were on the bench, then he would be Irish"? I think that a person would articulate his uneasiness in the following way:

> Granted that everyone on the bench is in fact Irish, it doesn't follow that their being Irish *is a result* of their being on the bench.

This is certainly correct, but notice that a new locution has found its way into the discussion, namely, "is a result of." If we insist on expanding every subjunctive conditional of the form "If x were ϕ, then x would be θ," into "If x were ϕ, then, as a result, x would be θ," we will certainly resist the move from a generalization of accidental correlation to a singular subjunctive conditional. But there is absolutely no reason to interpret subjunctive conditionals in this fashion, just as there is no reason to demand that every justificatory argument should admit of the same expansion. "Since x is ϕ, x is θ" can not always be expanded to "Since x is ϕ, as a result, x is θ."

We might take a closer look at this locution "as a result of." The phrase naturally suggests some kind of causal relationship, and in many cases this is precisely what we intend. It is important, however, to realize that this locution can signal non-causal relationships. A square is a quadrilateral as the result of being a closed plane figure bounded by four straight lines, and here the relationship is definitional. An obligation can be the result of making a promise, and in this case, the relationship is grounded in the performative use of language. It is hard to find a common thread that runs through all of these uses of the phrase "as a result," and on top of this, each of these uses is a locus of philosophical controversy. Even so, there are some modest points to be made that will help us on our way.

It will help to have a label for inferential constructions that admit of an *as-a-result* expansion: we can call them consequential inferential constructions. The generalization "Everyone on the bench is Irish" warrants the inferential construction "If x is on the bench, then x is Irish," but it does not warrant the consequential inferential construct "If x is on the bench, as a result, x is Irish." On the other hand, the generalization "Anyone born in Ireland of Irish parents is Irish" certifies the consequential construction:

> If x is born in Ireland of Irish parents, then, as a result, x is Irish.

Here the consequential relationship is in some way definitional.

Relying on this rough and ready notion of consequential constructions, we can note two things that will help us to understand their logical characteristics:

> (1) We should not expect the instantiation of a non-consequential construction to have a consequential character.

I expect that the person who claims that a singular subjunctive conditional can not be derived from a generalization of accidental correlation is gratuitously giving the subjunctive conditional a consequential reading, and then, of course, the derivation appears to fail. The second principle is this:

(2) In general, the consequential character of an inferential construction is not preserved under transposition.

Thus the generalization "When it rains, the ground gets wet" supports the consequential construction:

> If it rains at time t, then, as a result, the ground gets wet at time t.

A simple application of transposition yields this unhappy result:

> If the ground does not get wet at time t, then, as a result, it is not raining at time t.[15]

But sanity is restored if we delete the consequential component from the inferential construction:

> If the ground does not get wet at time t, then it is not raining at time t.

This asymmetry between cause and effect is but a special case of the asymmetry between antecedent and consequent in a consequential construction.

The same principal holds for the transposition of a singular inferential construction: we can safely transpose such a construction provided that we delete the consequential component. We can not, of course, pass from $(p^+ \rightarrow q^+)$ to $(\sim q^+ \rightarrow \sim p^+)$, for here there is an incompatibility of epistemic commitments. In order to allow for transposition, a provision must be made that will conserve epistemic commitments. These conservation principles are given in the following table:

p	$\sim p$
p^+	$\sim p^-$
p^0	$\sim p^0$
p^-	$\sim p^+$

Using this table, we can now write out each of these constructions with its corresponding transponent:

I	$(p^+ \rightarrow q^+)$	$(\sim q^- \rightarrow \sim p^-)$	VI
II	$(p^0 \rightarrow q^+)$	$(\sim q^- \rightarrow \sim p^0)$	V
III	$(p^0 \rightarrow q^0)$	$(\sim q^0 \rightarrow \sim p^0)$	III
IV	$(p^- \rightarrow q^+)$	$(\sim q^- \rightarrow \sim p^+)$	IV
V	$(p^- \rightarrow q^0)$	$(\sim q^0 \rightarrow \sim p^+)$	II
VI	$(p^- \rightarrow q^-)$	$(\sim q^+ \rightarrow \sim p^+)$	I

I claim that all these transpositions hold provided that we respect the rule that transposition does not, in general, preserve the consequential character of an inferential construction.

VI

The discussion of consequential constructions leads naturally to the topic of explanations (and ultimately to an examination of the speculative use of subjunctive conditionals). The first point that I wish to insist upon is that in giving an explanation we are making an assertion and not offering an argument. This is a rather small point and can be brought out by comparing these specimen sentences:

(1) Since the floodgate jammed, as a result, the dam burst.
(2) The dam burst as a result of the floodgate jamming.
(3) The dam burst because the floodgate jammed.

The first sentence formulates an argument with its consequential character made explicit. The second sentence states that such a relationship holds. Thus (1) presents the basis for the specific sort of explanation formulated in (2). And (3) also formulates an explanation but with a less specific indication of its basis.

In using this example I have not tacitly accepted the covering law model of explanation. I think that it is clear that arguments of a consequential character can serve as the basis for *one sort* of explanation. It is not at all clear that every explanation must have as its basis this kind of argument, and for that matter, the basis for an explanation may not be an argument of any kind.[16] But if we restrict our attention to explanations that do have a consequential inferential construction for their basis—we can call them C-explanations—we can apply our reflections to problems that have persisted in the literature.

What is the relationship between C-explanation and prediction? The first step in answering this question is to drop the talk about prediction since it introduces irrelevant and bothersome problems about temporal relationships. Instead, we can ask about the relationship between C-explanation and justification. Stated in this manner, the answer is easy enough: when an explanation has as its basis an argument of the kind here discussed, then that same argument could also be used for the justification of its conclusion. To this we might add that when a C-explanation has as its basis a statistical argument, the strength of the C-explanation is directly proportional to the strength of the underlying argument when used in a justificatory manner.

---

[15] This construction makes sense under the epistemic reading: "If the ground does not get wet at time t, then as a result *we know* that it is not raining at time t." As a matter of fact, a further expansion is needed: "If we *know* that the ground is not wet at time t, then as a result, we *know* that it is not raining at time t." Of course, this follows from the original only if it too is given an epistemic reading. It should go without saying that endless confusion arises if we make such shifts without realizing that we are doing so.

[16] For a discussion of explanations with non-argumentative bases see William Dray, *Laws and Explanations in History* (London, Oxford University Press, 1957).

It is also obvious that a sound justificatory argument may lack consequential character and thus be unsuitable as the basis for a C-explanation. For example, we can account for the length of a shadow through citing the height of the building that casts it, but we cannot account for the height of the building by citing the length of the shadow it casts. The length of the shadow is, at least in part, the result of the height of the building, but the height of the building is in no way the result of the length of its shadow.

Thus if we concentrate upon C-explanations, two things are clear:

(1) The basis of a C-explanation provides the basis for a justification as well.

(2) Not every justificatory argument provides the basis for a C-explanation.

If we leave the domain of C-explanations, the relationship between justification and explanation becomes more obscure, especially if we consider explanations whose bases are not argumentative in structure.

VII

Thus far we have examined only the refutory function of subjunctive conditionals, now, through ideas introduced in the last section, we can investigate the other important way in which subjunctive conditionals function, namely, in the formulation of *speculations* concerning what *would have happened* if things had been different. We can start with a specimen sentence:

> If Cleopatra had been grotesquely ugly, she would not have had a child by Caesar.

Since we are speculating about what would have happened had things been different, we naturally give this construction a consequential reading:

> If Cleopatra had been grotesquely ugly, as a result, she would not have had a child by Caesar.

This example provides another opportunity to test the conjecture that, in general, transposition will not preserve consequential character. Under the assumption that we are dealing with a construction of type VI, $(p^{-}\rightarrow\sim q^{-})$, we can transpose it into a construction of type I, $(q^{+}\rightarrow\sim p^{+})$. But if we fail to delete the consequential element, we get the following unhappy result:

> Since Cleopatra had a child by Caesar, as a result, she was not grotesquely ugly.

Unless we appeal to the desperate principle that a woman is never so beautiful as when she is pregnant, it would be hard to imagine how this construction could be certified. These troubles disappear once we drop the consequential element from the construction:

> Since Cleopatra had a child by Caesar, she was not grotesquely ugly.

I think that it is clear that this expression holds if the original does, but it is also worth noting that a person engaged in speculation may be utterly uninterested in it since he does not wish to *prove* anything about the past.

The discussion of the speculative use of subjunctive conditionals raises another point of grammar. In speculating we seem to use subjunctive conditionals not only when we take the antecedent to be false, but also when we take it to be problematic. Suppose that I do not know if Jones is in his office. Here I might speculate in the following way:

> If Jones were in his office, I could ask him for a raise.

Since I am merely speculating, and not seriously considering the action, I do not hand myself an inference ticket. If I were seriously contemplating the action, I would put matters in the following way:

> If Jones is in his office, I can ask him for a raise.

When speculating, we are not interested in establishing the truth or falsehood of the component propositions, and it is for this reason, it seems, that we avoid inference ticket language.

VIII

With these ideas in mind, we can attack the two difficult problems concerning subjunctive conditionals: the first concerns the so-called problem of cotenability; the second centers around troubles that arise when we speculate from a counterfactual hypothesis. I shall discuss the first problem in this section, the second problem in the next section.

The best way to introduce the problem of cotenability is to turn to the writings of Nelson Goodman. Starting rather in the middle of things, we can pick up Goodman's argument at the point where he offers the following tentative suggestion for necessary and sufficient conditions for the truth of a counterfactual:

> . . . a counterfactual is true if and only if there is some set S of true sentences such that S is compatible with C and ~C, and such that A.S is self compatible and leads by law to C; while there is no set S′ compatible with C and with ~C, and such that A.S′ is self-compatible and leads by law to ~C.[17]

Here *A* is the antecedent of the conditional, *C* the consequent of the conditional, and *S* introduces the further conditions needed to carry out the deduction from the antecedent to the consequent.

Goodman's next move is to declare that these requirements are not strong enough:

> As an example, consider the familiar case where for a given match, we would affirm
> (i) If match M had been scratched, it would have lighted,
> but deny
> (ii) If M had been scratched, it would not have been dry.
> According to our tentative criterion, statement (ii) would be quite as true as statement (i). For in the case of (ii), we may take as an element in our S the true sentence
> Match M did not light,
> which is presumably compatible with A (otherwise nothing would be required along with A to reach the opposite as the consequent of the true counterfactual statement (i). As our total A.S we may have
> Match M is scratched. It does not light. It is well made. Oxygen enough is present . . . etc.
> And from this, by means of a legitimate general law we can infer
> It was not dry.
> And there would seem to be no suitable set of sentences S; such that A.S′ leads by law to the negate of this consequent. Hence the unwanted counterfactual is established in accord with our rule.[18]

Having stated the problem, we can next turn to the diagnosis of it.

> The trouble is caused by including in our S a true statement which although compatible with A would not be true if A were. Accordingly we must exclude such statements from the set of relevant conditions; S, in addition to satisfying the other requirements already laid down, must be not merely compatible with A but "jointly tenable" or *cotenable* with A. A is cotenable with S, and the conjunction A.S is self-cotenable, if it is not the case that S would not be true if A were.[19]

Goodman's argument ends with a coda that seems to leave matters in a hopeless state:

> . . . in order to determine whether or not a given S is cotenable with A, we have to determine whether or not the counterfactual "If A were true, then S would not be true" is itself true. . . . In other words, to establish any counterfactual, it seems that we first have to determine the truth of another. If so, we can never explain a counterfactual except in terms of others, so that the problem of counterfactuals must remain unsolved.[20]

Wilfrid Sellars has pointed out that Goodman's argument has this surprising feature: it is nowhere shown in detail how the problem is generated by the demand for cotenability. Sellars further argues that Goodman has misdiagnosed the problem and that once the problem is correctly understood, a solution to it can be found. Since the Sellars' argument is complex, dealing with a variety of problems at once, I shall not attempt to repeat it here. Instead, I shall try to show how the notions developed in this paper augment his approach to the problem.[21]

We can begin with Goodman's first specimen counterfactual:

(i) If *M* had been scratched, it would have lighted.

There is no difficulty in expanding this construction in a way that exhibits its consequential character:

(i′) If *M* had been scratched, as a result, it would have lighted.

---

[17] Nelson Goodman, *Fact, Fiction, and Forecast* (London, The Athlone Press, 1954), p. 20.

[18] *Ibid.*, p. 21.

[19] *Ibid.*, pp. 21–22.

[20] *Ibid.*, p. 23.

[21] Wilfrid Sellars, "Counterfactuals, Dispositions, and the Causal Modalities," *Minnesota Studies in the Philosophy of Science*, Vol. II (Minneapolis, University of Minnesota Press, 1958). Sellars has also been kind enough to offer detailed criticisms of this essay and has thus saved me from a number of embarrassing mistakes.

On the other hand,

(ii) If *M* had been scratched, it would not have been dry.

does not admit of an expansion to:

(ii′) If *M* had been scratched, as a result, it would not have been dry.

No point of logic blocks this expansion, for we can imagine a world where matches get wet as the result of being scratched. This gives us the *first* clue to the problem: perhaps part of the difficulty is that in the setting of Goodman's paper we are tempted to give both these constructions a consequential interpretation when in fact only the first construction will bear this interpretation. These reflections lead to a further question: Can we legitimately give (i) a consequential reading without thereby being committed to a consequential reading of (ii)? I shall now show that the answer to this question is *yes*, and therefore that the two constructions are not logically on a par, as Goodman thinks we are forced to concede.

We begin by constructing a somewhat complex covering law. In the vernacular, the law would take the following form:

If a match is scratched, as a result, it will light, provided that it is dry, well made, etc.

The thrust of Sellars' argument is that we must pay close attention to the individual roles played by the expressions within this covering law. In particular, we must avoid any willy-nilly translation of this law into PM-ese. Following his suggestions, but not his exact procedures, I will employ four different symbols to flag the various components within the covering law:

K for a kind of thing
C for standing conditions (items under the proviso clause)
A for the antecedent of a consequential construction
R for the consequent of a consequential construction

I will adopt the following symbolism for the items in this covering law:

Mx x is a match
Dx x is dry (we will ignore the remaining conditions)
Sx x is scratched
Lx x lights.

These symbols and the appropriate flags yield the following translation of the covering law:

$$Mx_{K} \rightarrow (Dx_{C} \rightarrow (Sx_{A} \rightarrow Lx_{R}))$$

As a first exercise, we can derive the subjunctive conditional (i′) from this covering law together with the following assumptions:

I	II
Mm	$\sim Sm$
Dm	$\sim Lm$

These items are divided into two categories because they enter into the derivation in two different ways. Using only the items in the first category, we arrive at the following result:

$$(Sm_{A} \rightarrow Lm_{R})$$

Now using the items in column II, we can ascribe the following epistemic commitments:

$$(Sm_{\bar{A}} \rightarrow Lm_{\bar{R}})$$

Since this has the flagging of a consequential construction, we may legitimately read this:

(i′) If *M* had been scratched, as a result, it would have lighted.

All is well.

We can now attempt to derive (ii′) from the covering law together with the very same set of assumptions:

I	II
Mm	Dm
$\sim Lm$	$\sim Sm$

Notice that while the assumptions are the same, they enter into the derivation in a different manner. First using the items in column I we derive:

$$(Sm_{A} \rightarrow \sim Dm_{C})$$

Then in virtue of the items in column II:

$$(Sm_{\bar{A}} \rightarrow \sim Dm_{\bar{C}})$$

Everything is in logical order, but this time the construction does not have the flagging that warrants a consequential reading of the construction. This bears out the contention that given the same covering law together with the same set of further assumptions we can justify a consequential reading of (i) but not of (ii).

A further problem remains to be solved. Even if we refuse to give (ii) a consequential reading and thereby avoid one possible confusion, there still seems to be something decidedly odd about the remark:

(ii) If *M* had been scratched, it would not have been dry.

By way of experiment, we can transpose (ii) into a construction of type I:

Since the match was dry, it was not scratched. Even if we read this as a justificatory argument and thereby protect ourselves against an illicit consequential reading, the construction still seems peculiar. Why? The trouble, I think, is that information has been suppressed in a highly unorthodox fashion. Notice that the following construction does not seem peculiar:

> Since the match was dry and *did not light*, it was not scratched.

If we wish to signal more clearly that an argumentative move has been made, we can express this in the following way:

> Since the match was dry and did not light, it *could not have been* scratched.

Once the expression symbolized by "$\sim Lx$" is made explicit, the oddness disappears, and this is a reflection of the fact that we often suppress standing conditions, but only under somewhat unusual cases do we suppress the other items in the covering law. If we now return to (ii) we can make it appear sensible by the very same device:

> If M had been scratched and had not lighted, then it could not have been dry.

Once more the special phrasing in the consequent brings out the justificatory force.

To sum up, we can say that the oddness of Goodman's unwanted counterfactual is derived from two sources: the first is our tendency to give it a consequential reading that it will not bear, the second is the unorthodox manner in which it suppresses information. Once we compensate for these two features, the construction no longer seems odd, and we are not embarrassed by the fact that it can be derived from the same set of assumptions that warrants a sensible counterfactual.

IX

The final topic of this paper is a set of problems that arise when we speculate from a counterfactual hypothesis. We can start with Quine's famous example:

> If Bizet and Verdi had been compatriots, Bizet would have been Italian.
> If Bizet and Verdi had been compatriots, Verdi would have been French.

How are we to decide which, if either, of these constructions is acceptable?

We might spell out these arguments in a little more detail:

Assume: Bizet and Verdi were compatriots.
Facts: (1) Compatriots are of the same nationality.
(2) Bizet was French.
(3) Verdi was Italian.

The assumption together with Facts (1) and (2) certify the first construction; the assumption together with Facts (1) and (3) certify the second construction.

Nicholas Rescher has suggested a way out of this difficulty.[22] In order to reason consistently from the given counterfactual assumption at least one of the items under the heading of Facts must be excluded from the reasoning. Well, which item should go? Rescher's important insight is that this is an extra-logical question, for each of the true items has an equal claim upon our assent. In fact, *the content of our speculation remains indeterminate* until we dialectically specify which facts may enter into our reasoning. Once the parameters of the speculation are established *on extra-logical grounds*, the reasoning itself moves along standard logical lines.[23]

X

The point of the last two sections is to clear subjunctive conditionals of any *special* culpability. There are problems concerning subjunctive conditionals, but these problems inhere in the entire system of inferential constructions of which subjunctive conditionals form only a part. An outstanding remaining problem is to specify necessary and sufficient conditions (if there be such) for a *consequential* construction. I suspect that this is the implicit goal that led Goodman and others to turn their attention to subjunctive conditionals. If this suspicion is correct, we must then admit that Goodman's basic problem remains unsolved. Nonetheless, the problem has now been freed from the special setting in which Goodman posed it, and this gives us a clearer conception of the place of this problem on the logical map.

Pomona College

Received November 4, 1965

---

[22] Nicholas Rescher, *Hypothetical Reasoning* (Amsterdam, North Holland Publishing Company, 1964). I have also benefited from Rescher's comments and criticisms of this essay.

[23] In passing we might note that a similar problem does not arise when a subjunctive conditional is given a refutory function. In refuting something we need only show that the antecedent together with *any* set of true premisses leads to a false conclusion.

American Philosophical Quarterly
Volume 4, Number 1, January 1967

III. THE CHOICE BETWEEN UTILITARIANISMS

DAVID BRAYBROOKE*

According to act-utilitarianism, there are no possible circumstances in which the principle of utility, applied to single actions with full consideration of the features of the circumstances contributed by previous actions and the influence which the actions now contemplated may have on future actions, ought to be set aside in deference to particular moral rules. According to rule-utilitarianism, there are such circumstances: circumstances may appear in which particular moral rules ought to be upheld, even though the harm done by single violations, taking them as single instances, independently of actions done independently from them, would be outweighed by the good that the violations would achieve. The rules must themselves be justified by utilitarian considerations; but they are rules of a type that sometimes may require counter-utilitarian actions.[1]

Of these views, rule-utilitarianism is the one more difficult to state without paradox; yet it is also the view that matches traditional morality more closely. Does it involve "rule-worship"? The support that it gains from traditional morality cannot be decisive for accepting rule-utilitarianism as a normative theory. Utilitarianism in the Benthamite tradition has always meant to be both normative and innovating. But rule-utilitarianism can be innovating, too: the rules that it justifies may not be traditional ones. Moreover, the inflexible attachment that it requires to some of the rules that it justifies is not irrational, given certain political conditions, that is to say, given certain arrangements (or lack of arrangements) for communication and moral legislation. As it happens, these conditions are like those that traditional morality has had to deal with.

There are other arrangements—other social forms within which moral policies may be developed—in which rational grounds for applying rule-utilitarianism in preference to act-utilitarianism disappear.

The Community-in-Session. Imagine a society small enough so that all its members can be gathered in one arena. There, in public, the members spend their whole lives, or at any rate all those tracts of time in which they might do anything that might awaken moral concern in any other member. No one can escape publicity. If two people should turn their backs on the rest of the community, they find themselves facing mirrors visible throughout the arena. If they whisper together, their words are picked up by a microphone and broadcast.

In such a society, one might infer, it would be sensible for anyone proposing to do anything that was in any way *prima facie* objectionable to seek community approval first. Let us suppose that, in fact, it is a universal custom for everyone in the community to do this, so that every decision about which even minor moral doubts might be raised becomes a public decision in which the whole community takes part. Let us suppose further that every such decision takes the form of sustaining or modifying existing moral rules; or of legislating new ones. The society may thus be regarded as a legislature continuously in session. Let us ascribe to it full sovereign powers, and call it "The Community-in-Session."

---

* Professor H. Chandler Davis of the University of Toronto, Professor Harry R. Pitt of the University of Nottingham, and Dr. Richmond H. Thomason, then a student in the Yale graduate school, have very kindly assisted in the preparation of this paper by providing substantial advice both mathematical and philosophical. An earlier version of the paper was read at the University of Rochester on March 5, 1963.

[1] "Act-utilitarianism" and "rule-utilitarianism" are Brandt's terms. See Richard B. Brandt, *Ethical Theory* (Englewood Cliffs, N.J., 1959), ch. 15. I follow Smart in adopting them. (See J. J. C. Smart, "Extreme Utilitarianism: A Reply to M. A. Kaplan," *Ethics*, vol. 71 (1960–61), p. 133; and *An Outline of a System of Utilitarian Ethics* (Melbourne, 1961), pp. 2, 4–5. I do not subscribe to Brandt's definition of rule-utilitarianism (*op. cit.*), pp. 396–397. The characterization that I give above (pointed for the purposes of this article) is compatible with his definition, but not equivalent to it. The case for rule-utilitarianism (as characterized above) does not, it seems to me, presuppose that there is a *complete* set of ideal prescriptions which is to be adopted on utilitarian merits exceeding those of other sets. (There is, so far as I can see, no trace of this presupposition in Smart's treatment either.) I treat act-utilitarianism and rule-utilitarianism as doctrines about how the principle of utility is to be applied if thorough-going application of it is sought. Rule-utilitarianism contends that act-utilitarian application fails in some possible cases (though not in all).

Besides providing against concealment, the Community-in-Session provides, let us suppose, against the formation of factions or coalitions that seek to promote some members' selfish interests at the expense of others. The provisions against concealment will themselves operate against factions. We can go further by assuming—since we need to assume a rule for voting anyway—that the community follows the rule of unanimity in voting.

These are not yet sufficient provisions. Having the unanimity rule will not by itself remove temptations to break it. If there are two groups in the community, one with one color of skin and one with another, or two groups that have different native tongues, or two groups that have different religious traditions, there may be danger of one group being mobilized against the other, regardless of adverse publicity. So we shall assume that the community is sufficiently *homogeneous* to forestall dangers of this kind.

We shall assume further that every member of the Community-in-Session rates his chances of surviving *arbitrarily* formed coalitions without being victimized as too small to be worth acting on. (As examples of arbitrarily formed coalitions, consider the series: (1) the 600 people on this side of the fountain vs. the 500 on the other; (2) the 350 people in the sun vs. the 250 in the shade; (3) the 200 standing up vs. the 150 sitting down; and so on.)

Finally—simply to avoid having to take up a certain set of problems—let us suppose that the Community-in-Session is so fortunate as to escape any military threats or natural emergencies substantial enough to require sacrificing the lives or fortunes of some members as a means of preserving the lives or fortunes of the rest.

Leaping over the difficulties of elaborating the principle of utility,[2] I shall assume that agents do frequently know what the net undistributed benefits of given actions will be; and that when they know what the net undistributed benefits will be they can always rank those actions according to the amount of those benefits (whether or not they can measure amounts exactly). I shall also assume that when the net undistributed benefits to be obtained from an action are known to be greater than those to be obtained from alternative actions, either distribution will not be called for or there will be a distribution to all the present members of the community such that every member gets what all members regard as a fair share. The breathtaking scope of these assumptions is a measure of the importance of the problems about meaning and justice which they sweep over.

Act-Utilitarianism in the Community-in-Session. The version of utilitarianism that would prevail in the Community-in-Session (if any did) is act-utilitarianism. If there were no other political arrangements to consider, there would be no point even in keeping up the distinction between act-utilitarianism and rule-utilitarianism; for in the Community-in-Session the grounds for preferring rule-utilitarianism disappear.

In the Community-in-Session, no member can ever choose to act secretly. Brandt's would-be gas-cheat[3] and Smart's (earlier, A. K. Stout's) flower-gardener,[4] if they wished to except themselves from the existing rules about using gas and water, would have to try to make out cases that would persuade the rest of the community. They would have to show that their using more gas or more water than other people would imply a net increase in happiness (or in the means of pleasure) if other people did not follow suit. They would also have to show that the knowledge of exceptions being made for them would not have adverse consequences—such as, other people in fact following suit; or growing refusal on many people's part to heed any common policy.

In the Community-in-Session, no member is ever in a position where he must act on his own, if he is to act at all. Every member is always in communication with every other member, and every action is an immediate possible subject of renewed legislation. No one is ever closeted with a dying testator; no one makes promises on desert islands. If anyone believes that departing from an existing rule will be beneficial in a case that unexpectedly arises, he can seek to have the rule amended. *The absence of personal discretion is compensated for by continuous and immediate access to the*

---

[2] For the purposes of the present discussion, I follow Brandt and Smart in waiving inquiries into the difficulties. In a book that C. E. Lindblom and I have written together, *A Strategy of Decision: Policy Evaluation as a Social Process* (New York, 1963), the difficulties are faced and coped with in a way that dispenses with a felicific calculus and bypasses the difficulties of Paretian welfare economics.

[3] Richard B. Brandt, *Ethical Theory* (*op. cit.*), pp. 389–391.

[4] J. J. C. Smart, "Extreme and Restricted Utilitarianism," *The Philosophical Quarterly*, vol. 6 (1956), pp. 344–354; at pp. 350–351.

CALCUTTA CENTRAL LIBRARY

moral authority of the community (which is not to say that authority is infallible).

If everyone in the Community-in-Session adopted utilitarianism, and thereby agreed on the same unambiguous criterion for choosing actions and policies, then every single action taken in the community would conform to the principle of utility (apart from mistakes due to ignorance). Every action will confer greater net benefits on all the members of the community affected by them than the alternative actions contemplated.

Rule-Utilitarianism in Intermittent Communities. Let us now imagine that the same people who have been living together as the Community-in-Session disperse throughout the area of a New England town or of a small Swiss canton. They will no longer be continuously ready to legislate moral rules. They will meet as a legislature only intermittently, perhaps once a day or once a week, perhaps only once a year or once a decade. However frequently the community meets, there will now be intervals between sessions during which members of the community may have to act on their own in morally significant, but doubtful cases. These are circumstances that open up a genuine option between act-utilitarianism and rule-utilitarianism.

Imperfect Agents. The most obvious reasons for preferring rule-utilitarianism in intermittent communities derive from the imperfections of men. Men are, notoriously, imperfect in moral attitude and in knowledge and intelligence as well.[5]

It would be very helpful, no doubt, if every agent would make only those decisions which he could anticipate the community's ratifying—when and if the community ever meets in legislative session, and supposing that only utilitarian arguments would be publicly persuasive when it did. But by the common opinion that men have of themselves, this cannot be counted on. A community in which this cannot be counted on may have a use for customs or institutions that are intended to limit—in crucial cases, to remove—the discretion of individual agents when they are out of communication with the community as a whole. If the dangers of letting them act on their own discretion are greater than the dangers of binding them, it will conform to the principle of utility to bind agents by general rules established in advance. People will be expected, for example, to keep promises as a matter of rule; and the point of having a rule will be precisely to rule out personal discretion about keeping or breaking promises in individual cases.[6]

However, possibly the common opinion of mankind exaggerates the imperfections of agents' characters. Experiments might show that if people are enjoined to be act-utilitarians and are trusted with discretion in carrying out the injunction, they will not in most cases abuse the discretion. It may turn out, at any rate, that they will be less tempted to abuse discretion if it is given them than they would be to flout rules which allow them no discretion, but which (in crucial cases) cannot be enforced except by internalization in the agents' consciences. This is an empirical question about variations in character.

There is another empirical question to consider: besides contemplating the possibility of agents' characters varying, temptations remaining the same, one may contemplate the possibility of temptations varying, agents' characters being held constant. Communities find themselves in distinctive historical environments. In certain environments, given certain courses of events, it is possible that a policy of giving agents full discretion to act as act-utilitarians would give better results in the long run than binding them not to use such discretion—even if common suspicions about the proportion of agents who would abuse discretion are fully vindicated. For the aggregate gain in happiness from the cases in which discretion is properly used may outweigh the aggregate loss from the cases in which it is not. In this respect also rule-utilitarianism, so far as concerns the obvious reasons for preferring it cited above, rests on empirical foundations that are by no means perfectly settled for all times and places.

---

[5] As a sophisticated champion of act-utilitarianism, Smart gives these imperfections a great deal of weight. His act-utilitarians are superior to ordinary men in being rational, benevolent, and conscientious; and even they will do best to act by "stereotyped rules" in many circumstances. Cf., for instance, *An Outline of a System of Utilitarian Ethics* (*op. cit.*), pp. 4, 29–31.

[6] Cf. John Rawls, "Two Concepts of Rules," *The Philosophical Review*, vol. 64 (1955), pp. 3–32; at p. 16. The moral rules that will figure in my discussion may or may not have the same content as rules of traditional morality; and they may or may not be clothed with the extensive institutional vestments that Rawls, discussing traditional rules, has ascribed to "rules of practices." On the other hand, they are more than "summary rules," if these are understood as merely indicating the tendency of past experience as defeasible guides to present personal decisions. In all cases, they are designed to forestall the operation of personal discretion. As I aim to show, the difference between act-utilitarians and rule-utilitarians is not that act-utilitarians have no inflexible rules; the difference is that act-utilitarians must be supposed to enjoy fuller opportunities for legislation.

They are nonetheless without doubt the foundations that traditional morality has accepted. Traditional morality has given more weight to the dangers that discretion would be abused than to the advantages of a system of complete discretion. This weight explains the superior attractions, from the point of view of traditional morality, of rule-utilitarianism as compared with act-utilitarianism (though neither traditional moralists nor rule-utilitarians might want to say that rule-utilitarianism does explicate traditional morality).

The accepted foundations go further than this. They explain why the attachment to certain moral rules may appear blind or irrational or intuitive. Such attachment may be just the attitude toward moral rules which it is best to inculcate to bind people efficaciously when they are out of sight.

However, the foundations, even if they were settled, as well as accepted, would not justify this attitude in the eyes of rule-utilitarians, if the attitude prevented changing bad rules or obsolete rules. True rule-utilitarianism cannot be somnolent rule-utilitarianism. If rule-utilitarians are to live up to their own professions regarding the principle of utility, they must be alert to defects in existing rules. They must also bear in mind the fact that even the most judiciously selected rules are liable to become obsolete. For example, the human imperfections on which the necessity of binding rules supposedly rests are liable to change. If individual agents should become, on the average, more knowledgeable or benevolent, the rules will probably need adjustment to allow agents larger spheres of discretion.

Part of the power of arguments like Smart's for preferring act-utilitarianism lies in the fact that alertness to changes in men and environments seems more readily associated with act-utilitarianism than with rule-utilitarianism. In fact, both require alertness. In intermittent communities, differences in the speed with which people holding the two doctrines respond to perceived changes in conditions could be removed to any degree of approximation by making the intervals between legislative sessions shorter and shorter.

Quasi-Perfect Agents. The case for preferring rule-utilitarianism to act-utilitarianism in intermittent communities gains strength from the assumptions of traditional morality regarding human imperfections; but the case does not depend on those assumptions. Suppose that all the members of the community can be trusted to act according to the principle of utility. Suppose that they are all equally well-informed, and such accurate reasoners that they never make mistakes deductive or inductive. Since they are not, we suppose, omniscient, they are not perfect agents; but they are at least quasi-perfect—paragons of virtue and wisdom. Yet if the community to which these paragons belong meets at most only intermittently, they may have to accept limits on their discretion that will prevent them from acting as act-utilitarians in every instance. The principle of utility may require them to act as rule-utilitarians.

Suppose there is a sort of actions A that will in the aggregate violate the principle of utility if (but only if) n_h or more of them are done during a certain interval of time. The quantity n_h thus represents the minimum number of actions A during that interval such that V, the total net benefit from doing such actions, is negative. Suppose that the figure n is known and known to be greater than one. Suppose also that it is known that during the interval n_h (or more) agents will be in a position to do an action A.

Now, this may be the case in every single instance: If the agent considers the actions of every sort, including actions of the sort A, that he knows have occurred previously, and considers the actions of every sort, including actions of the sort A, that would be influenced by his doing action A now, he will on the basis of these considerations find himself enjoined to do A by the principle of utility; for doing A in each single instance, so considered, offers a certain benefit, a, greater than the perceived disadvantages.

But suppose that the community will not come into legislative session at any time during the interval. Suppose also that no set of agents s in number, where s is less than n_h, will be able during the interval to set up arrangements for communicating to the rest of the community that they are doing actions A less than n_h times in time to forestall $(n_h - s)$ agents from doing such actions also.

What does utilitarianism prescribe in this situation? The answer to this question turns out to be exceedingly complex. What utilitarianism prescribes varies according to whether the situation is discovered before or after the interval begins and according to what further information is in the possession of the agents and the community. It also varies according to whether or not everyone is expected to reason in the same way and act on the same principles (even though their final actions may differ); for the time being, let us assume that they are expected to. Beginning with the further assumptions that the situation regarding actions A is discovered by individual agents during the interval and that their information about the situation is limited to the facts set forth above, let us consider some cases.

(i) *Discovery during the interval: virtual legislation*

Under the conditions assumed, if every agent reasons and acts as an act-utilitarian, n_h (or more) actions A will be done during the interval and aggregate harm will result. It therefore contravenes the principle of utility for every agent to reason and act as an act-utilitarian. For there is another line of reasoning which would forestall the harm if every agent adopted it. According to this line of reasoning, the proper thing for every agent to do is to impose on himself the requirement of refraining from doing actions A, since otherwise every agent with the opportunity to do A will do A on act-utilitarian grounds. If doing A involves breaking a previously legislated rule, the self-imposed requirement amounts to reaffirming the rule. If doing A does not involve breaking a rule, the universal coincidence of self-imposed requirements may be regarded as virtual legislation. In both cases, the reasoning of the agents (which a utilitarian observer would endorse) is rule-utilitarian, although in the second case the rule has not been communicated, and symmetry of reasoning and conduct takes the place of actually legislating and heeding legislation.

On the present assumptions, Brandt's gas-cheat and Smart's flower-gardener, finding themselves in a community of quasi-perfect utilitarians, and yet subject to their familiar temptations, would have to refrain from the actions which they contemplate. For if they found the actions utilitarian by reasoning from certain considerations, so, they must recognize, would every other agent in a position to do the same thing; but if every agent in a position to do these things—to steal gas or use extra water—were to do them, the community would in each case suffer a net loss in utility. There is no way of preventing this result except by everyone considering himself bound to refrain. Therefore, everyone would refrain.

The fact that in these examples the attractions of A depend upon A's being a deliberately concealed action is not essential to justifying rule-utilitarianism in the type of situation assumed. What counts in compelling agents to adopt rule-utilitarianism in all such situations is the incompleteness of information and the want of communication respecting actions A, whatever the reasons for these deficiencies may be.

The effect of the present assumptions in other examples treated by philosophers is problematic. Would actions breaking promises in those instances where (taking the instances singly) breaking the promises gained more utility than keeping them aggregate into a harmful body of actions? It seems far from obvious that in *any* interval of time the number of such actions would be likely to reach a critical figure n_h. It is even less clear that the subclass of this class of actions which consists of *secret* infidelities would reach the critical figure. If all the deathbed agreements and desert-island promises that were ever made were broken in those instances in which doing so maximized utility, would the aggregate effect be counter-utilitarian? I find this difficult to believe.

(ii) *Discovery before the interval: legislation beforehand*

The facts about actions A may be discovered before the interval begins, in time for the community to meet and legislate. If the community then knew (in addition to the information given above) which number of actions, n_v, would maximize V, the total net benefit from doing actions A, the principle of utility would require legislating a rule according to which a number of agents would be selected to do A n_v times (and no more) during the forthcoming interval while all other agents were strictly forbidden to do A at all. Would it be act-utilitarianism or rule-utilitarianism that prevailed during the interval? (The collective action of adopting the rule would be act-utilitarian, provided that the rule itself does not introduce new disadvantages which did not figure among those attendant upon doing or failing to do A.)

The rule itself would be inflexible. Furthermore, every agent with an opportunity to do A would still be in a position, when the opportunity occurred, to reason that neither the actions A (or actions of other sorts) which he knows individual agents have done previously nor the actions A (or actions of other sorts) which would be influenced by his doing A make the benefits of his doing A less than (or even equal to) the disadvantages. The rule therefore contravenes act-utilitarian reasoning carried to this point. On the other hand, *given the rule*—and the legislation as a previous collective action that must be taken into account, the disadvantages of doing A will outweigh the benefits, except for designated agents doing actions A within their allowance. Carried to the end, therefore, the reasoning of every agent may be said to be act-utilitarian. This reasoning is legitimate only because of the acknowledged existence of an inflexible rule. But abiding by the rule will (on the assumptions made) be act-utilitarian in every instance.

If n_v were not known, the community might still be able to arrange to have that number of actions *A* done (and no more). It might be able to set up a communication center for use during the interval, legislating beforehand that every agent who did an action *A* was to report doing it to the center and every agent considering doing such an action was to obtain from the center an up-to-date compilation of other agents' reports. Then actions *A* would be done, one by one, up to the number of n_v such actions each justified by act-utilitarian reasoning; but on the point of doing an (n_v+1)th action *A*, any agent concerned would refrain. As a quasi-perfect agent in possession (thanks to the communication center) of the requisite information, he would see that, given all those actions *A* already done, the harm of doing one more would outweigh the good (and would offset in part the net total benefit, *V*, previously achieved).

During the interval, every agent will reason as an act-utilitarian about actions *A*, and the fact that it is possible for him to justify doing *A* by such reasoning only so long as every agent is inflexibly following the rule of reporting to the communication center and seeking information from it makes no difference to the character of his reasoning about actions *A*. But what about his reasoning concerning this rule? Here (by contrast with a rule designating n_v agents beforehand) it is by no means obvious that failing to abide by the rule will forfeit some net benefit available to the community or do the community harm. If the number of actions *A* reported to the communication center has not in fact yet reached n_v, what harm does an agent do if he disregards how many actions *A* have occurred? Since the compilation that he would obtain from the center would not deter him from doing *A*, what does it matter in his case if he omits to obtain the compilation?

Clearly, an agent would do harm by failing to report doing *A* when he does it. Failing to report would infallibly lead to more than n_v actions being done by all agents together, so that some agents would be doing net harm. Abiding by that part of the rule is act-utilitarian in every instance. But the part requiring that information be sought as well as given will rest on another foundation. As regards this part, rule-utilitarianism will prevail. If some agents assume that it is safe to disregard this part of the rule, all agents will assume this (since everyone reasons in the same way). But if all agents act on the basis of this apparent safety, more than n_v actions will be done—in fact, n_h or more. The only way to prevent this result is for all agents to abide by the second part of the rule. Therefore all agents will abide by it.

Suppose, finally, that though the situation regarding actions *A* is discovered before the interval begins, n_v is not known, yet circumstances preclude the community from setting up a communication center and proceeding as just described. The problem, again, is to modify the situation by legislation so as either to forestall act-utilitarian reasoning during the interval or to supply such reasoning with new premisses. What sort of rule should the community legislate? It might legislate a rule forbidding every agent doing *A*. Then rule-utilitarianism will prevail during the interval. In every single instance the attractions of doing *A* remain unchanged, but if any agent acts upon the attractions, all will. The only way available to prevent the aggregate harm consequent upon n_h (or more) agents doing *A* is for all to abide by the rule and refrain. Therefore all will reason that they must refrain.

Would it not be possible, however, for the community to obtain some of the *benefits* of actions *A* without running the risk that so many such actions would be done as to result in aggregate harm? Suppose the community legislates that some number of agents are to be designated to do actions *A* less than n_h times, while other agents are to be forbidden to do *A* at all. Then, though *V*, the total net benefit from actions *A* during the interval, will not necessarily be maximized, it will at least be greater than zero. Any rule that had this result would conform better to the principle of utility than a rule forbidding every agent doing *A*. Moreover, experience with situations of the sort assumed might enable the community to stipulate a number of actions *A* to be done that would have a good chance of being close to n_v: for example, on the basis of experience, it might stipulate fifty per cent of n_h.

What sort of reasoning during the interval would support abiding by a rule of this sort? Abiding by the rule could, of course, be loosely justified on rule-utilitarian grounds by comparing the effect of everyone's abiding with the aggregate harm that would occur if no rule of any sort existed to limit the number of actions *A* to be done. But this is only loose justification (as it would be for other rules that we have considered), since it does not by itself exclude agents from making act-utilitarian attempts to improve on the rule during the interval. Would such attempts be forthcoming and could the rule be defended against them?

The answer, it seems to me, depends upon how confident the agents are that the stipulated number of actions A is likely to be close to n_V. The more confident they are that it is likely to be close, the firmer are the act-utilitarian grounds for abiding by the rule in every instance. To deviate will more likely than not diminish V. The less confident they are, the more scope there is for rule-utilitarian reasoning.

If the stipulated number is thought to be a wild guess likely to be much smaller than n_V, any agent not designated to do A might reason on act-utilitarian grounds that he should nevertheless do it, since doing A beyond the stipulated number will increase V. But if one such agent reasoned this way, all would, and all would act accordingly. To prevent this from happening, all will abide by the rule. On the other hand, if the stipulated number is thought to be a wild guess likely to be much greater than n_V, any agent designated to do A might reason on act-utilitarian grounds that he should *refrain* from doing A as required by the rule. But if one such agent reasoned this way, all would, and all would refrain, with the result that V would diminish to zero. To prevent this from happening, all will abide by the rule.

If the stipulated number is thought to be a wild guess, but no one knows what the direction of the error is likely to be, then neither those agents who are to do A by the rule nor those who are to refrain can firmly justify on act-utilitarian grounds either abiding by the rule or flouting it. The loose rule-utilitarian justification mentioned above will remain, however, and in the absence of other considerations it would prevail.

Smart's Mixed Strategy Solution. Let us revert to the assumption that the situation regarding actions A is discovered during the interval, while the agents are out of communication with one another. One can hardly help being tantalized by the thought, is there no way, even with this assumption in force, of co-ordinating the decisions of those quasi-perfect agents so that some actions A are done, in order to reap the acknowledged benefits, but fewer than n_h, so that general harm is avoided? Better yet, is there no way of finding and co-ordinating on the n_V actions that would maximize the total benefit V (not merely prevent it from becoming negative)? Smart has called in game theory to treat the gas-cheat and flower-gardener cases, with results that may come as close to finding n_V and co-ordinating on n_V actions, when the agents are out of communication with one another, as any approach to the problem is likely to get.

Smart argues that in a society composed entirely of act-utilitarians, individual agents should follow *mixed strategies* with regard to actions of the sort confronted in the gas-cheat and flower-gardener cases. They should make use of appropriate chance devices to give themselves a certain probability p of doing the action and another probability $(1-p)$ of refraining. Thus

> Let m be the number of people in the community. Let $f(n)$ be the national damage done by exactly n people disobeying the government's request [not to use excess gas; or not to water gardens]; it will be an increasing function of n. Now if each member of the community gives himself a probability p of disobeying the edict it is easy to determine, as functions of p, the probabilities $p_1, p_2 \ldots p_m$ of exactly 1, 2, ... m persons respectively disobeying the edict. Let a be the personal benefit to each person of disobeying the edict. I am, of course, supposing what is perhaps a fiction, that numerical values can be given to $f(n)$ and to a. Then if V is the total benefit to the community we have:
> $$V - p_1(a-f(1)) + p_2(2a-f(2)) + p_3(3a-f(3)) + \ldots p_m(ma-f(m)).$$
> If we know the function $f(n)$ we can calculate the value of p for which $\frac{\mathrm{d}V}{\mathrm{d}p} = 0$. This will give the value of p which maximizes V.[7]

This argument does open up new possibilities of co-ordination that are more discriminating in effect than the universally prohibitory rule previously discussed. However, the solution to which the argument leads is not quite so attractive from a utilitarian standpoint as it seems at first sight. It certainly seems at first sight that if all agents in the community calculate p and follow the mixed strategy embodying it, that V, the total benefit to the community, will be maximized; that, in other words, n, the number of people disobeying the edict (each once) will equal n_V, the number whose disobedience will maximize the total extra benefit to the community of doing any of these acts of disobedience at all (provided at least that a is reinterpreted to include benefits not only to the agent disobeying but also to other people). It seemed so to me, and it would have continued to seem so if I had not taken counsel with a mathematical friend (H. Chandler Davis). His advice enables me to point out that what Smart's solution would maximize is something very different from V, the total benefit, which is a simple function of the number of people disobeying. Smart's solution would maximize E, the expectation of benefit, that is to say, $\sum_{n=0}^{n=m} p_n V(n)$. This is the quantity, not V, that is described by the right-hand side of Smart's equation.

---

[7] J. J. C. Smart, *An Outline of a System of Utilitarian Ethics* (*op. cit.*), pp. 43–44. It is assumed that V, the total (increased) benefit, equals 0 when $n = 0$; and that V is not then maximized.

Quite consistently with the agents undertaking to behave so as to maximize $\sum_{n=0}^{n=m} p_n V(n)$, a chance may remain, perhaps a substantial chance, of so many agents being in fact led to disobey that V, far from being maximized, would turn negative. It may be granted that the chances of V's becoming drastically negative—a disaster—would be small.[8] Even so, cautious agents may prefer the no-risk policy of inflexibly obeying the edict.[9] They may be especially inclined to prefer this policy if there are many sorts of actions A_1, A_2, A_3, etc., all of them avoidable, that would form so many different routes to disaster if a sufficient number were done of any of them. The probabilities of engendering disaster by these different routes may add up to an over-all risk large enough to deter the agents from taking the mixed strategies approach to the actions class by class.

These complications are not necessarily unmanageable. However, they are not the only difficulties in accepting Smart's argument, illuminating as it is, as vindicating act-utilitarianism even for agents who are out of communication with one another. In the first place, Smart's assumptions are unrealistic (as he himself freely admits). They are several degrees more unrealistic than the assumptions that we worked with above, which led to a rule-utilitarian solution when the situation regarding actions A is discovered during the interval. Smart assumes not merely that a, the benefit to be expected from each action taken singly, is known, but that it is the same amount in the case of every agent. The latter assumption is perhaps no more heroic than assuming that a varies between agents (which is by itself more plausible) but that at least some agents know all its values. At least some would have to, if any were to apply Smart's argument.

The assumption that some agents know the whole function $f(n)$ redoubles the heroism. The most that people are likely to know, in any real situation, are some values of n greater than n_h, which have even stronger implications of disutility, and some values of n at the other end of the range, very much less than n_h. If they expect n_h agents (or more) to be in a position to do actions of the sort in question during a given interval,

---

[8] Richmond H. Thomason has supplied me with the following illustration of a mixed strategy solution, which is typical in indicating how the chances of a disaster remain but are minimized:

"Consider a community of three individuals. Here,

$$E(p) = \sum_{i=0}^{3} \binom{3}{i} p^i (1-p)^i (a.i + f(i))$$
$$= (1-p)^3 f(0) + 3p(1-p)^2 (a+f(1)) + 3p^2(1-p)(2a+f(2)) + p^3(3a+f(3)).$$

I have added the values of f in this equation, since I am thinking of f as a function which gives the utility to the community, positive or negative, of a certain number of people performing a given act. In the cases which interest us, f will be nonpositive and decreasing (or at least nonincreasing).

Let $a = 1$, and let f be given by $f(0) = 0$, $f(1) = -\frac{1}{9}$, $f(2) = -\frac{2}{15}$, and $f(3) = -7$.

One can imagine situations which would give something like these numbers; say, three homeowners, each with electric dishwashers. If one or two of them use their appliance in a given time period (say, after the dinner hour) there is only a little general inconvenience, say, a slight dimming of the lights, but if all three wash at once, the fuse blows on the main line, and someone has to go out in the cold to replace it.

The value of p which maximizes expected utility (i.e., $E(p)$) here is $\frac{1}{4}$.

If the three individuals spin their wheels, in an average run of 64 days none will wash 27 days, 1 will wash on 27 days, 2 will wash on 9 days, and all 3 will wash (and consequently blow the fuse) on 1 day. Total utility gained in this period, using Smart's method, will be

$$27.0 + 27.(1-\tfrac{1}{9}) + 9(2-\tfrac{2}{15}) + 1(-4), \text{ or } 24 + \tfrac{84}{5} - 4 = 36\tfrac{4}{5}."$$

[9] V represents utility; hence it already takes into account the impact that various gains and losses would have on the community, given its initial circumstances. A loss of a given size will mean more, one may assume, to a poorer community than to a richer one. Agents in either community may still vary among themselves in the utility or disutility that they place on taking even favorable risks. A positive value of E may be offset in a given agent's eyes by the disutility to him of gambling. See Rudolf Carnap, *Logical Foundations of Probability* (Chicago, University of Chicago Press, 1950), p. 276.

tempted to do them by various amounts of prospective benefits a, a rule-utilitarian solution emerges, for they cannot undertake Smart's calculations. To be consistent utilitarians they will all have to refrain from doing A.

In the second place, Smart's mixed strategy solution is itself a rule-utilitarian solution.[10] For it does envisage circumstances in which the principle of utility, as applied to single instances, ought to be set aside and actions done that are counter-utilitarian, taken as single instances. This will happen whenever the chance device dictates that the agent should refrain from doing A. The (virtually legislated) rule that is upheld in such circumstances may be identified as the same rule upheld when the chance device dictates that A is to be done, namely, the rule that the mixed strategy incorporating p is to be followed. It is true (as Smart points out) that this rule is quite unexpectedly different in kind from traditional moral rules. Nevertheless it is not a rule founded on the principle of utility as applied to single instances; it is not a rule liable to variation from instance to instance, considering in each instance actions known to have occurred previously and actions expected to occur subsequently as a result of the one now contemplated. It is a rule founded not merely on those considerations, but also on the application of the principle of utility to aggregate effects, including the effects of actions taken independently, parallel to one another but otherwise unconnected.[11]

Improving on Smart's Solution by Prior Legislation. If the situation regarding actions A is discovered beforehand and the community thus has the opportunity to legislate a rule to take care of the situation, Smart's solution will be discarded. For then, since n_V can be calculated, the community will (as demonstrated above) adopt a rule designating a number of agents to do A n_V times during the interval, forbidding other agents to do A at all.[12] This solution (again, as remarked above) would be act-utilitarian throughout: the rule would be adopted on act-utilitarian grounds and abiding by the rule during the interval would in every instance be justified by act-utilitarian reasoning.

Mixtures of Reasoners. Hitherto it has been assumed that every agent would reason in the same way. In circumstances in which act-utilitarianism sufficed, all agents would be act-utilitarians; in other circumstances, for reasons compelling to them all, all would be rule-utilitarians. Among the latter circumstances, however, there are some in which though the principle of utility would be better served by all agents being rule-utilitarians than it would be by all agents being act-utilitarians, it would be still better served if some agents were the one and other agents the other. Consider again the circumstances in which the situation regarding actions A is discovered during the interval, without sufficient information coming to light for Smart's mixed strategy solution to be applied. If in these circumstances it is known that $(n_h - 1)$ or fewer of the agents belonging to the community are convinced act-utilitarians, while the rest are convinced rule-utilitarians, more net benefit will be obtained for the community if everyone acts according to these different convictions than will be obtained if all agents are converted to rule-utilitarianism. In the former case, some positive amount of V will be obtained from actions A; while in the latter case, everyone would refrain from doing A, and hence V would be zero.

---

[10] This point has been established in the course of an exchange of views between Smart and Morton A. Kaplan in *Ethics*. See Kaplan, "Some Problems of the Extreme Utilitarian Position," *Ethics*, vol. 70 (1959–60), pp. 228–252; Smart, "Extreme Utilitarianism: A Reply to M. A. Kaplan," *ibid*, vol. 71 (1960–61), p. 133; and also Kaplan, "Restricted Utilitarianism," *ibid*., pp. 301–302.

[11] Kaplan rightly stresses the importance of the concession that Smart makes to him about the assumption of symmetry, namely, that in assuming symmetry of reasoning and conduct a departure has been made from act (extreme) utilitarianism. For in assuming symmetry, Smart takes aggregate effects fully into account; whereas act-utilitarianism takes aggregate effects into account only insofar as they are connected with given single actions—either as known from past actions, and in this way entering the premisses of this one, or as expected among the causal consequences of the one now contemplated. "The rightness or wrongness of an action is to be judged by the consequences, good and bad, of the action itself" (*Outline*, p. 4). "The rightness or wrongness of keeping a promise on a particular occasion depends only on the goodness or badness of the consequences of keeping or of breaking the promise on that particular occasion" (*Philosophical Quarterly* article, p. 344). Act-utilitarianism, so expressed, is inconsistent with a mixed strategy solution. Smart wavers between recognizing the discrepancy (in his reply to Kaplan) and refusing to recognize it (in the *Outline*, p. 44). In the *Outline*, Smart assumes that the members of the community in which the mixed strategy solution is adopted are all act-utilitarians; but they cannot be. They may have begun by being act-utilitarians; but they become rule-utilitarians as soon as they adopt the mixed strategy solution. "Rule-utilitarianism holds that the rightness or wrongness of an action is to be judged by the goodness or badness of the consequences of a rule that everyone should perform the action in like circumstances" (*Outline*, p. 4). The rightness or wrongness of doing A, whatever its attractions *per se*, is to be judged by the consequences of a rule that everyone should perform the action when and only when the chance device, set for the solved value of p, signals "Go ahead." Is Smart an act-utilitarian or a rule-utilitarian? The answer is, he is an act-utilitarian given certain circumstances; and a rule-utilitarian given others. That is exactly what a consistent utilitarian ought to be.

[12] Commenting on the illustration given above, Thomason says, "If the three individuals are in communication beforehand and work out an agreement so that two can wash on every night of the period in the same time (the time of an average run of 64 days), $64(2-\frac{2}{15}) = 119\frac{7}{15}$ units of utility will be gained, a considerable improvement over the expected utility to be gained by the mixed strategy solution, which was $36\frac{4}{5}$."

In such circumstances, Brandt's gas-cheat and Smart's flower-gardener, taking them to be act-utilitarians in a mixed community of act-utilitarians and rule-utilitarians, might go ahead and do A in perfect accord with the principle of utility, except insofar as the expectation of their acting contrary to rule might undermine the morale of the rule-utilitarian segment of the community or make social training in ethics less effective.

The exception, of course, is a big one when it applies. On the point of morale, it does not apply if we preserve our assumption that when the net benefits of an action or policy are greater than the net benefits of the alternative actions or policies contemplated, the benefits will be distributed so that everyone gets what all regard as a fair share. But suppose that the benefits redound exclusively to the act-utilitarians. Then, even if the rule-utilitarians do not suffer any concrete harm, it might well seem to them that the benefits obtained by the act-utilitarians were unjustly obtained; without the uncompensated self-denial of the rule-utilitarians the act-utilitarians could not act. Moreover, whether or not the rule-utilitarians have this feeling of injustice, would it not damage any system for social training in ethics to discard or flout the premiss that everyone ought to have the same ethics? It may seem obvious that ethical training must be the same for everyone.

Yet the circumstances described do suggest that, contrary to the assumptions of philosophers and of traditional morality,[13] uniformity in ethics is not an unmixed good. Moreover, in the circumstances described, though the rule-utilitarians may be under a natural temptation to resent the permission that the act-utilitarians have of doing A, would it not be unworthy of them (and their utilitarian principles) to deny that a better social result will in fact be obtained by diversity of reasoning and conduct? Are they to be dogs in the manger? Furthermore, one might imagine that the special position of the act-utilitarians was not only not resented, but respected. The act-utilitarians might all be retired war heroes, or film stars, or philosophers, granted a franchise to practice act-utilitarianism as people deserving just such a tribute and indulgence.

The Old Man of the Sea. If the moral community meets to legislate sometime *during* the minimum interval within which the aggregate effect of agents' freely doing A would be harmful, it may meet in time to recognize that some number of actions A less than n_h have been done, accept without recrimination the good thereby accomplished, and flatly forbid further actions of the same kind until the interval ends (or by some other scheme of limitation, like those discussed earlier, prevent the number from rising to n_h before the end of the interval, and maybe from rising beyond n_v). Even if the situation regarding A is discovered before the interval begins, it might possibly make sense to postpone legislating about the situation until some part of the interval has elapsed.

But suppose, as before, that the community will meet to legislate only at the end of the interval. Then whatever sort of agents it consists of, imperfect or quasi-perfect, it may need a rule about actions A. Let us suppose that the rule which it is compelled to adopt, in actual or virtual legislation, is one that is supported by rule-utilitarian considerations, which in this connection defeat act-utilitarian considerations. Now, the case for any such rule will have to be an empirical one, given that the community is composed of utilitarians. But then, however well-reasoned, it may be a mistaken case. People may discover during the interval that there is no chance that n_h or more actions A will take place during the interval; or that there are no grounds for thinking that greater harm than good would eventuate if they did. What are these people, acting as individuals, to do?

Their responsibility will be somewhat lightened if the community will soon meet to legislate. Meeting at the end of the interval, the community will be able to weigh the utility gained by adhering to the rule against the utility forfeited by denying any exceptions; it may be expected to correct this and every other rule so far as they need to be modified, and usefully can be. But this is not an expectation that many people can have in the real world. The large-scale societies characteristic of the real world in modern times do not meet to take up moral legislation. In them, individual agents who have discovered the disutility of rules hitherto thought to be useful cannot be restrained by the promise that the community will legislate on the subject at the earliest opportunity. There are provisions of sorts for something like moral legislation in such societies—but the provisions are slow, multiple, and confused. Often they will not even be set in motion unless individual moral reformers take the initiative of presenting test-cases of violation.

---

[13] As analyzed, for example, in Marcus G. Singer's discussion of *Generalization in Ethics* (New York, 1961).

If an individual agent, discounting soberly for his own liability to error, has discovered that a certain inflexible rule is mistaken, and that violating it would do more good than harm, even when the scandal of violating it is allowed for, will he not be bound by the principle of utility to violate the rule? But then what, in the end, becomes of rule-utilitarianism? Is not act-utilitarianism finally vindicated? Evidently an agent is bound to heed a rule, however "inflexible" in formulation, only so long as acting in accordance with it maximizes utility in the case of every single action.

This conclusion does not follow. For the discovery that the agent makes, in the circumstances assumed, is not merely a discovery about the alternative actions immediately in prospect. It is also—and more significantly—a discovery about a rule, and about just one rule at that. To discover merely that violating the rule on this occasion would maximize utility, even taking into account the effects (if any) of the violation in this instance in encouraging violations in subsequent instances, would not upset the rule. The case for the rule and for the rule's being inflexible stands or falls not on these effects alone (which may be negligible), but on the aggregate harm of actions that may be taken quite independently. What the agent discovers, if he discovers that there is no case for the rule, is that this aggregate harm has been overestimated. His violation can be justified only by proof that there was no need for an inflexible rule in this connection. But that proof does not extend to showing that there is never any need for inflexible rules. It does not even extend to bringing into question the validity of such other inflexible rules as may be currently in force. The case for every inflexible rule has to be considered on its own merits.

Let it be noted, furthermore, that the agent's discovery is compatible with either of two very different conclusions respecting future actions of the sort *A* that have hitherto fallen under the discredited rule. It may be the case that in view of the agent's discovery, no inflexible rule can be justified for governing any future actions of this sort. On the other hand, it may be the case that one or more new inflexible rules can be justified as governing one or more subclasses of these actions. It may be better to allow promises to be broken whenever there are utilitarian grounds for breaking them than to have an inflexible rule against breaking promises on any occasion. It may be better still to have an inflexible rule restricted to governing promises (say) about money.

Here, as in other circumstances, utilitarianism reveals a challenge to the adequacy of our institutional means for establishing or ratifying fine discriminations in moral insight. But the position in pure theory is quite firm. Whatever its difficulties, rule-utilitarianism is not to be surrendered to act-utilitarianism at all times and all places. In some circumstances, depending on the state of the world and on the state of information, act-utilitarianism would displace rule-utilitarianism. In other circumstances, a consistent utilitarian will have to stand by rule-utilitarianism.[14]

Dalhousie University

Received October 29, 1965.

---

[14] David Lyons' *Forms and Limits of Utilitarianism* (Oxford, The Clarendon Press, 1965) has appeared since the present article was written. In the course of an argument whose intricacy will fascinate philosophers, Lyons undertakes to show that what he (rather disobligingly) calls "primitive rule-utilitarianism" is "extensionally equivalent" to act-utilitarianism, yielding "equivalent judgments in all cases" (p. x; cf. p. ix, p. 25). It turns out that this result depends chiefly on the contention that any discrepancy between reckoning the effects of acts separately, then adding the effects up, and reckoning their effects combined, when they occur in combination, can be removed by redescribing the single acts so as to apportion to each its full share of the combined effect. I think that this contention may be freely admitted. It does not persuade me that such discrepancies do not need to be kept in view, and an important distinction between "act-utilitarianism" and "rule-utilitarianism" drawn presupposing the possibility of such discrepancies. A utilitarian spectator imagining himself in possession of perfect information might have no use for such a distinction. Utilitarian agents liable to have imperfect information, imperfect opportunities for communication, and varying opportunities for legislation would find the distinction vital. I hope that I have helped to show why.

American Philosophical Quarterly
Volume 4, Number 1, January 1967

IV. PERFORMATIVES

HERBERT FINGARETTE

I

IN his most elaborate study of the performative use of language,[1] Austin mentioned certain crucial and intuitively obvious features of the performative, but he failed to appreciate them adequately. It is for this reason, I believe, that the initially distinctive notion seems to dissolve before his very eyes. He ends by "producing a programme" rather than proposing a definitive account of the nature of the performative. In the remarks which follow, I propose to take up again certain of the original and intuitively central features of the performative, and to refine and elaborate them enough to show that the notion, as originally (though obscurely) conceived, is a distinct and sound one. I shall turn back from the indirect or less sharply differentiating approaches ultimately adopted by Austin and other later commentators on the notion.[2] An analysis based on what is intuitively central will also reveal why and how this distinctive way of using language is peculiarly suitable for counting as the performance of certain important kinds of acts.

The first thing I wish to do is to present a brief and informal account of the performative by paraphrasing or quoting various remarks of Austin's and interweaving a few comments of my own. With only the briefest subsequent explicit review of Austin's reasons for not remaining within the framework of such an account, I shall then proceed to explore its possibilities systematically.

When I say, in the appropriate circumstances, "I apologize," or "I promise," I am not describing or reporting—saying something true or false about—what I am doing; in saying it I am doing it, i.e., I am apologizing or promising. Thus, "saying so makes it so." This motto helps to bring out something distinctive about the performative. But this motto, like any motto, can easily mislead unless carefully qualified.

How must "saying so makes it so" be qualified? When I promise, I may "say so" but I need not "say so"; that is, either "I promise to do it," or simply "I'll do it" may suffice. Further, while there is an obvious sense in which when I promise I may "say so"—i.e., by saying, "I promise"—I am not "saying so" in the sense that I am saying *that* something *is* so; I am not making a true or false assertion. Finally, "saying so" is not in and of itself enough to "make" it so. The saying is usually what Austin calls a leading incident, often *the* leading incident, in the performance of the act. But the saying is not the sole thing necessary—not any utterance of "I promise" by anyone at any time under any circumstances and regardless of frame of mind constitutes making a genuine promise. Broadly (and vaguely) speaking, the circumstances must be right or appropriate, the use must be a "happy" one.

Nevertheless, we must not forget that although the performative utterance may fail to "make it so," for the utterance is not the sole thing necessary, yet in the appropriate circumstances it is, so to speak, the case that the utterance cannot fail. That is, to name the ship *is* to say (in the appropriate circumstances) the words, "I name, etc." The force of this is that there is a peculiar authority, efficacy, or necessity in the way "saying so makes it so," for it seems that what I "make so" when I utter the words in the appropriate circumstances is not merely that I have uttered words, but also and infallibly the fact that the ship is so named.

Austin first concentrates his original attention upon the circumstances which must be "right" if a performative is to be "happy." He spells out the general nature of these conditions (his A, B, and C conditions),[3] and then discusses them.

---

[1] J. L. Austin, *How to Do Things With Words* (London, Oxford University Press, 1962).

[2] Since this paper is primarily a constructive exposition of the nature of performativeness, I shall limit my critical remarks to Austin, and briefly to Sesonske's paper (see below) which brings out a point in which I am specially interested. Among papers I have in mind here, however, are: W. O. Alston, "Linguistic Acts," *American Philosophical Quarterly*, vol. 1 (1964), pp. 138–146; A. Sesonske, "Performatives," *The Journal of Philosophy*, vol. 62 (1965), pp. 459–468; P. F. Strawson, "Intention and Convention in Speech Acts," *The Philosophical Review*, vol. 73 (1964), pp. 439–460.

[3] Austin, *op. cit.*, pp. 14–15.

Then comes Austin's first major proposal for distinguishing the performative from the constative (e.g., "he is running"). We might say, remarks Austin,[4] that it is the fact that he is running which makes the constative utterance "He is running" true; whereas it is the happiness of the performative "I apologize" which makes it the fact that I am apologizing. This proposal seems to Austin to break down when he discovers that "considerations of the happiness and unhappiness type may infect statements (or some statements) and considerations of the type of truth and falsity may infect performatives (or some performatives)."[5]

It is by the time he has reached this juncture that Austin has committed, in my opinion, two fundamental mistakes which thereafter lead him further and further away from the heart of the matter.

Mistake No. 1: Austin tacitly assumes that the performative and constative uses of language must be mutually exclusive. Therefore the fact that considerations of the happiness type "infect" constatives, and that considerations of the true-false type "infect" performatives, is evidence to him that these two types of considerations cannot suffice as the basis for distinguishing performative from constative. This unnecessary assumption of mutual exclusiveness may have been inspired by Austin's bias at this earlier stage of his analysis toward finding what he calls "grammatical" or "lexicographical" criteria of performativeness. Sesonske has pointed this out and has argued (correctly, I believe) that we would do better to consider the distinction essentially a functional one.[6] It is evident that the same "grammatical" form may be used for different purposes simultaneously. And indeed Austin later on notes in passing that "some verbs may be used . . . simultaneously in two ways . . . (embracing) both a performative utterance and a description. . . ."[7] But his argument in general ignores the larger significance of this isolated comment.

We can pinpoint the place at which the crucial misstep occurs. At the very place Austin accepts as decisive the "infection" argument noted above, he discusses only one example to support his thesis, and his discussion is extremely brief. He uses as his single example of a performative: "I warn you that the bull is about to charge" (said when in fact the bull is not about to charge). This performative statement, he says, is subject to the charge (characteristic of the constative) that it is false or mistaken, but cannot be accused of failing "in any of the ways we have hitherto characterized as varieties of unhappiness." We may admit that Austin is right in saying that there is something false or mistaken about this statement; but contrary to what Austin says, it is *also* "unhappy." For his condition A2, essential for "happiness," requires that the "circumstances in a given case must be appropriate for the invocation of the particular procedure invoked."[8] And typically the circumstances are not appropriate for the warning procedure when in fact no danger impends. Austin's tacit assumption of mutual exclusiveness requires that the statement be either constative or performative. But if we allow that the same sentence may be used performatively and constatively simultaneously, then we can easily see that "I warn you that the bull is about to charge" is constative insofar as it says something false, i.e., that the bull is about to charge; and, it is an unhappy performative insofar as it is intended as a warning but is uttered in an appropriate circumstance (of type A2).

Mistake No. 2: Although Mistake No. 1 is important and must be corrected if we are to understand the performative, correcting that mistake is by no means enough. For at best we still remain with the proposal (Austin's second mistake) that the criterion of performative function is the relevance of the happiness tests. Unfortunately the happiness tests are also relevant to utterances which it is intuitively obvious have no performative function but are paradigms of the constative. Examples are: "John's children are all bald," said when in fact John has no children; my saying "The cat is on the mat" when in fact I don't believe this to be true. Even with the correction of Mistake No. 1, we must finally admit, as does Austin, that his "happiness" tests do not suffice to distinguish performative from constative.

Austin's tests fail to suffice for identifying the performative because he has lost sight of the one "condition" which is most characteristic of the performative. My first and central thesis concerns the nature of this most characteristic feature of performativeness. The thesis, roughly put, is that in the performative use of language the central burden of one's remark expressly *concerns the fact that one is making the remark*. Or, to put this point the other way around: it is the performance of the act *per se*, the making of the remark, which is *central to the message* of the remark.

---

[4] *Ibid.*, p. 47.

[5] *Ibid.*, p. 55.

[6] Sesonske, *op. cit.*

[7] Austin, *op. cit.*, pp. 64–65.

[8] *Ibid.*, p. 15.

One may think of the matter this way (as a preliminary): When everything else is ready and right for apology, the remaining object of central interest is that I actually *say* the words. There is some leeway as to what words I shall say: I may say, "I apologize," or "I'm sorry," or "I know it was unfair of me, and I'll never do it again." But whatever I choose to say, *I* must *say* it. Of course it is true that whenever I say anything, I must say it. But this "of course" is inappropriate for the case of the performative. For in this case it might be said that you wait for the gesture more than for the words. It is misleading, when I have apologized, to say that "of course" I "also" performed an act of speech—for in fact I was then and there performing that act. If that is not what I was doing, then it wasn't an apology at all. The fact of my then and there uttering certain words is of the essence of my message.

To a good extent I have been laboring an obvious point. Yet it needs laboring because, oddly enough, though it is both central and obvious, it has disappeared as such from all the attempts at an explicit and definitive characterization of the performative with which I am acquainted. I propose now to bring into sharper focus the elements of the problem and to develop a precise formulation of the nature of the performative.

II

The element of performativeness to which I first want to call attention is the special efficacy or authority of the performative kind of "saying so"—that authority by virtue of which saying so makes it so. In saying, "I apologize" (in the appropriate circumstances), the relation between my apology and what I say is clearly of a more intimate and necessary kind than that between the fact of his running and my statement, "He is running."

The closeness of this tie between performative use of language and accomplishment, its necessary rather than contingent character, is seen more sharply by considering an example which seems superficially to have this "saying so makes it so" character. The example I have in mind is: "I am now talking loudly." If I am talking loudly when I use this sentence, then the sentence is true, and in a way (but, as we shall see, not in the right way) my saying so made it so. For if I had not uttered the words as I did, it would be false that I was then talking loudly; and the fact that I did utter the words as I did was precisely what made my assertion true. And yet: it is quite evident that I may say, "I am now talking loudly" and my use of the sentence may be happy, yet be false. For I may say it in writing, or in Indian sign-language, or simply speak it softly. Since "saying so" does not require loud talking, my "saying so" does not of itself *necessarily* make it so, even if the circumstances are otherwise appropriate.

On the other hand, when I say "I apologize" in appropriate circumstances, it *is* so, "in the very nature of things." One can plausibly say here that the saying *is* the apology. This "is" expresses a necessary connection of some kind but not an identity. For after all, the utterance alone *is not* an apology: as we have noted, there are all sorts of ways in which it may fail to be an apology at all, and there are often a number of other utterances which would serve as well. There is clearly an essential element of contingency here too. Let us see if we can separate and describe precisely the various components of this cluster of necessary and contingent relationships. The lead I shall follow is that of the statement, "I am now talking loudly." This is a kind of saying which can make what is said so, and it does this in part by virtue of what we might loosely and for the moment speak of as "self-reference." If only we could add the element of necessity. Is there any kind of "self-referring" statement which, in addition, incorporates some relevant kind of necessity? Putting the question this way calls to mind that paradigm of "self-reference," the use of the expression "I". In the typical correct use of "I", the expression is used to refer to something which at that moment necessarily exists. Here we have an interesting partial analogy to a kind of necessity found in the performative "I apologize." It is not that it is necessary that I should now exist or that I should now be apologizing, but it is the case that *if* I am correctly using "I" or "I apologize" in certain typical ways, *then* it is necessarily the case that at that moment I do exist, or that I am apologizing.

This kind of necessity needs to be explored, but it will not do to remain too long with "I" as the center of interest. For the correct typical use of "I" does not *make* me exist in the quasi-causal way that my saying "I apologize" can make it the case that I am apologizing. My existence is a necessary *pre*-condition of my using "I", but it would surely be wrong to say that my apology is a precondition of my saying "I apologize." My saying "I apologize" *brings about* what was not so before. My using "I" does not bring it about that I exist. The "making" element, i.e., the bringing-it-about-that, is characteristic of the performative. Therefore I think it would help to avoid confusion and dilution of a very useful notion if we refused to follow Hintikka's lead in speaking of performativeness in connection with "I".[9] (I think that this point reinforces von Leyden's criticism that what is sound in Hintikka's results could be achieved without introducing the notion of performativeness.[10])

As it turns out, however, the Cartesian example is only a specially important instance from a larger class of uses of language where there is a necessary connection between the happy use of an expression with intent to refer and the existence of the intended referent. Furthermore, in some of these cases (though not in the case of "I") the necessarily existent referent is *brought into* existence by the use of the expression. In these cases, which I shall shortly discuss, it is appropriate to say that "saying so makes it so," where "makes" has the force of "necessarily brings it about that." What is this larger class of language uses within which fall both performatives and the typical use of "I"?

To make use of language in any particular instance, we must perform an action; we must carry on such activities as talking, writing, or gesturing. Although any of several actions may be adequate for what we may call essentially similar uses of expressions or sentences in a language, performance of some act or other is essential. The expression or sentence and its use are thus distinguishable from the particular action. For example: I can use the sentence "I intend to participate in the Commencement exercises" to inform you that I intend to participate, and I make this very same use of it regardless of whether my performance consists in writing on a note-pad and putting it in the office mail, or in talking to you, or in checking a questionnaire-sheet in the box which says, "Yes" rather than the one which says, "No." What I am here calling the use is the same; and the expression used is the same; but the activity in which I engage is quite different. Nevertheless I must engage in *some* activity or other—this is necessarily so.

The necessity in question is of a kind which is neither contingent nor merely a matter of definitions or "mere words." To understand this necessity is to understand something essential about the entire enterprise of using language.

From this standpoint we can readily specify certain necessary conditions for the use of linguistic expressions. One necessity is that there be a language-user, the speaker (if we allow the latter term to do duty for "talker," "writer," "gesturer," etc.). The use of an expression presupposes a context, and more particularly a time and a place of use. And most relevant to our purposes there is, necessarily and uniquely related to each instance of the use of an expression, its utterance. By "utterance" I mean that which the speaker does which is sufficient to make specific what expression he is using and what use he is making of it, the time and place of the doing being the time and place of that instance of the use of the expression. Note that the utterance may be in the form of the activity of writing, talking, gesturing, etc. Note also, therefore, that the unique and necessary relation of utterance to use is not such as to imply that only this utterance could have served for use of the expression. Instead, it is the case normally that any of several or many different utterances could have served, but *some one* of these is required, and the one which in fact is used to satisfy this requirement is the one which I call the necessary and uniquely related utterance.

I shall speak of the totality of such necessary conditions, as well as of *their* necessary conditions, as the "speech-act situation." And I shall speak of each such condition as a necessary element of the speech-act situation. My enumeration of necessary elements is not exhaustive. The speech-act situation includes everything that is necessarily so, given an instance of the use of a certain linguistic expression.

There arises the possibility, of special interest to us here, that a linguistic expression may be used to refer to its speech-act situation or to some necessary element thereof. The use of expressions to refer to (elements of) the speech-act situation as such I shall call *reflexive reference*. The notion of reflexive reference is not *ad hoc* but is of interest in more than one philosophical context. For example, reflexive reference is associated, as I have already suggested, with the Cartesian *cogito*. The normal use of the expression "I" is to refer to a necessary element of the speech-act situation, specifically, to whatever person is using the word. Thus we go from the fact of a correct and typical use of "I" to the conclusion that there is necessarily a speaker to whom it refers.

---

[9] J. Hintikka, "Cogito, Ergo Sum: Inference or Performance?," *The Philosophical Review*, vol. 71 (1962), pp. 3–32.

[10] W. von Leyden, "Cogito, Ergo Sum," *Proceedings of the Aristotelian Society*, vol. 63 (1962), pp. 67–82.

Of course the conclusion that there is a speaker is itself dependent upon the supposition that "I" has in fact been typically and correctly used. How do we determine this? Our assessment of the situation in this regard is a complex and not fully understood process, but certainly one important and relevant item in such an assessment is that we take it that there is a speaker. If we had genuine doubt whether there was a speaker, we would thereby have genuine doubt whether we were faced with a use of a linguistic expression rather than mere sounds of a language-like kind.[11]

Thus the existence of the speech-act situation and the possibility of reflexive reference generate a general form for a use of linguistic expressions for referring such that it is guaranteed that they *do* so refer. In all such uses there is an interesting combination of contingency and necessity—a kind of "synthetic *a priori*" character. For in this reflexively referring use of language, one makes a necessarily true claim about the world. The kind of claim I mean is the kind inherent in a referring use of an expression. "John's children . . ." in a typical use tacitly makes the claim that John has children—which may or may not be so. But "The words I am now uttering. . . ." and "here" also are used typically to make referential claims—that I am now uttering words, that I am at a particular place—and these claims are necessarily true.

Turning to the specific question of performatives, I can now present a precise version of my first and principal thesis that the burden of the performative remark concerns the fact that one is making the remark: The performative use of an expression is a use which reflexively refers to its utterance. In a more suggestive formulation for our purposes: the expression is used in a *performance-referring* way, i.e., to refer reflexively to the performance of an act (the utterance) by the speaker.

This formulation of my first thesis brings out an important part of that synthetic *a priori* character noted earlier in connection with performatives. The performance-referring use of the expression is a use where the speaker's deed, the performance *referred* to, *consists in* the speaking. It is thus the kind of *saying* so which in the appropriate way *makes* things so, i.e., it necessarily brings it about that what is said is so. We have here just that blend of necessity and of "causal" efficacy which characterizes the performative.

The characterization of the performative as the performance-referring use of an expression takes us a long way, but it does not take us all the way toward delineating a language-use whose crucial features are those we intuitively glimpsed as central to performativeness. Fundamental questions yet remain.

That I utter the words "I apologize" is intrinsically of little interest in itself; it is the apology which is of interest. However, my first thesis as stated above is not that we use the expression "I apologize" to refer to our apology but to the utterance. The utterance of itself is certainly not an apology: I might utter the same words, "I apologize," but do so as a demonstration of acting technique. Thus it is clear that more remains to be considered before we shall have shown how a performance-referring use of "I apologize" can constitute an apology (and before we shall have shown why such use of an expression is a suitable device for apologizing).

"I apologize" is a meaningful expression. Its meaning is a matter of the rules and conventions governing its use. Of course an important part of its meaning has to do with the practice of apologizing and with related practices, e.g., insulting, injuring, promising (and failing to keep one's promise). For it is in terms of these practices that the rules for use of "I apologize" are initially derived. The practice of apologizing includes, obviously, more than such things as *saying* "I apologize." It includes rules and conventions concerning the grounds for apologizing, the claims and commitments generated by apologizing and by the accepting of apologies, and so on. "I apologize" takes its meaning from the existence of this practice. To use "I apologize" is in this respect to *allude* to the practice. You cannot understand how I'm using "I apologize" unless you at least understand the significance or meaning of the expression. But if you do understand the expression, then in understanding me when I say "I apologize" you are grasping my allusion, by means of a use of that expression, to the practice of apologizing. I am not *referring* to the practice, but in using an expression with this meaning I am alluding to it. (Probably I am alluding to the practice whenever I say, "I apologize"—even when, for example, I am using the expression to demonstrate acting technique or to joke. But in these cases *further* convention has been invoked—play-acting, joking—which permits this allusion to the practice of apology while qualifying or excluding the exercise of the practice.)

---

[11] Problems arise which are interesting and important for a thorough exploration of the concept of the speech-act situation but which cannot (and need not) be explored in detail in a paper devoted primarily to the analysis of the performative. For example, what shall we say of the bottle DRINK ME which Alice found? Is there (necessarily) a speaker here? My answer would follow the line that there is a "speaker" necessarily presupposed insofar as we take DRINK ME to be akin to the typical use of "me"—and it is precisely out of this that there arises that personification which lends comic delight to tales such as Alice's. Problems also arise in connection with "the time and place of use" where we deal with labels or bottles, messages read much later than the time written, phonograph records, telephones, etc. Here again, a full analysis cannot be attempted, but it seems clear that the very question calls to our attention that "time and place of use" may be considered in some cases and for some purposes as referring to several times and/or places. It does not conflict with the assertion that there is always at least some time and place necessarily and uniquely related to each instance of the use of language. Finally, I might repeat that my list of the necessary elements of the speech-act situation is not exhaustive and omits discussion or even mention of a number of interesting items such as that the speaker is in some way conscious, and that there is a language which the utterance serves to use.

Thus, when I use performatively the expression "I apologize," I do so in a performance-referring way and also, at the same time, I allude to the practice of apology by virtue of the meaning of the expression. Now let us suppose that you understand the practice of apologizing and that you can (as you typically would be able to) understand the bearing of that practice in our present situation. Perhaps it is obvious that I might well apologize to you now. Perhaps it is not quite so obvious, yet to anyone (such as you) who understands the practice, it needs only an allusion to it at this moment for it to become obvious that its exercise by me is appropriate here and now. In saying in the present context, "I apologize," I explicitly allude to the practice and thereby to its appropriate exercise in this situation by me. The allusion is explicit; but it need not be. I might make the allusion but make it tacitly. For example, I might only mention, using a conventional tone and manner (an "apologetic manner"), some crucial ground for my apologizing: "I'm very unhappy about it," or "I shouldn't have done it; it was unfair of me." In the one case I use in the expression the very word which names the practice; in the others I do not. It is only in this not too important sense that the explicit-tacit distinction applies. For the use in either case is the same: the practice requires that a practice-alluding expression be used in a performance-referring way, and that this be done in a context where it is evident to the parties concerned that the practice alluded to is appropriately exercised by me then and there.

In an important range of practices—but by no means in all—it is the convention that such a combination of performance-referring and practice-alluding use of language counts *also* as the performance of a certain further act defined by that practice. Specifically: for me to say "I apologize" in the appropriate circumstances (where the practice of apology is relevant) and to do so in a performance-referring way, counts *also* (according to the conventions of apologizing) as the making of an apology by me then and there (i.e. at the time and place marked by my utterance).

An analogy may help to bring out more sharply the relation of the utterance to the act of apologizing. It is as if I were ostentatiously to drive into the ground a number of stakes and at the same time say, "Claim staking!" Driving the stakes is an action which corresponds here to uttering "I apologize"; it is intrinsically of no great interest, and of itself serves neither to fence in or otherwise protect my property. But by thus publicly singling out certain points in space and time, and by alluding in the very same act to the practice (here presumed relevant) of claim-staking, I have also, according to the relevant legal rulings, exercised that practice and, in particular, I have effectively staked a claim to the land.

We may now take note of features of the performative use of language (and of gesture) which add still further that peculiarly synthetic *a priori* character noted earlier. There are many ways in which surrounding circumstances may be inappropriate and make the use an unhappy one. For example, there may exist no such practice as I thought, or it may not be relevant, or my utterance or gesture may not count in the circumstances as any *further* act. Hence we have a distinct awareness of the *contingent* relation between the stake-driving or the utterance of "I apologize" on the one hand, and the staking of the claim or the apology on the other.

Yet, although this use of language in exercising a practice smacks of contingency, it is clear that the further act "brought about" is not brought about as causal consequence of the utterance. It is a purely "logical" consequence of an accepted rule or convention of the form "Doing A in circumstances B counts also as doing C," and the fact that "A has been done in circumstances B."

It is of special significance that the performance referring use of an expression is a particularly suitable device to serve as a marker, and the reasons for this special suitability deserve mention. Whoever understands my reflexively referring use of an expression must perforce have:

(a) observed an act, *A*, performed by me,

(b) understood me to have used a certain expression,

(c) understood that I am using the expression with the intent to refer to an act of mine,

(d) understood that it is necessarily the case that the act to which I mean to refer has in fact been performed by me (there is no possibility of error on my part or deception as to the actual occurrence of the act referred to),

(e) and understood with certainty which act is referred to: he knows for certain that the act to which I refer is necessarily the very one, *A*, which he observed for himself.

Furthermore, this very same expression can be simultaneously used to allude to the practice to which the marker is relevant.

Such a use of language to exercise a practice is peculiarly suitable for those cases where a person is directly, explicitly altering his responsibility with regard to others. There are, of course, other ways of altering our responsibility to others and theirs toward us. But in the performative we have an act which, in the ways I have outlined, removes the gap between word, intent, and deed, between what we say we do, what we mean to do, and what we do do. Intent is fused with Deed; Deed is transparent with Meaning. Such an act serves well for the most explicit, formal, and direct expression of will in establishing or altering claims, privileges, duties, rights. The performative is the act *par excellence* for explicit and formal commitment, for assuming, giving up, or rejecting responsibilities.[12]

This inherent aptness of the performative for altering "formal" relations has recently been emphasized by Sesonske,[13] though he offers no analysis of the mechanism by which the alteration is produced. He simply *defines* the performative as "an utterance whose point is to alter formal relations." Appreciation of this fact alone is a major development in our appreciation of performatives. However, Sesonske says that the use of language to alter formal relations is the defining property of performatives, whereas it is in fact, as I have noted above, an important task which performance-referring use of language is particularly suited to accomplish.

Of course Sesonske, as have others, at times takes note of what I have insisted is the central feature of performativeness. For example, Sesonske says, ". . . the saying of the words in the appropriate context constitutes the event." But the import of these remarks is not reflected finally in his definition. Without this element, his definition is, as we see, inadequate. But to include this element would require consideration of what is meant in this context by the expression "saying . . . constitutes the event." It is such expressions (and analogous ones used by Austin and others) which are central to performativeness and which my own account explicates.

At a time when I thought that my analysis up to this point was essentially complete, Professor Sesonske (for whose many helpful comments on this paper I am extremely grateful) challenged me to account for: "I do *not* apologize." This sentence can typically be used in a performance-referring, practice-alluding way in a context where apologizing would be appropriate. According to my account up to this point, "I do *not* apologize" apparently ought to be the performance of an apology, whereas it is in fact and evidently quite the contrary. Clearly it is necessary to specify why making this kind of remark does not, according to the practice, also constitute making an apology.

The performative use of language constitutes action, and doubly so, for it is, when successful, both the act of uttering words and the act of, e.g., apologizing. Where action is possible, alternatives exist. I may apologize, but then again I may prefer in those very same circumstances not to apologize. I may promise, but this presupposes that I may choose not to promise. Nor are the alternatives always only two. I may apologize or simply not apologize, or I may look you in the eye and say deliberately: "I do *not* apologize." This is different from simply not apologizing. To mention only one of several ways in which it differs: "I do *not* apologize" is itself a performative, a kind of commitment (though not absolute) to forego the privilege of apologizing and reject the rewards and costs of apology, whereas mere omission of an apology leaves it more open that one may yet exercise the privilege.

---

[12] A fuller discussion of the relation of commitment to responsibility, and thus in effect a large-scale elaboration of this important point about performatives, is to be found in my paper, "Responsibility," *Mind*, vol. 75 (1966), pp. 58–74.

[13] Sesonske, *op. cit.*, p. 467.

I do not always have, in the case of other practices, analogous options. Thus when the chairman of a meeting desires unanimous consent for some action, he may simply ask: "Are there any objections?" I can then object or not object, but there is no third option analogous to "I do *not* apologize." Furthermore, aside from these three sorts of alternatives, there are often also other sorts of alternatives, for example, half-apologizing (quite common), half-betting (impossible), and half-promising (a borderline case). Then too there are alternatives other than those along this yes–no dimension, alternatives of a multiple-choice kind. For example, there are the several kinds of alternative moves in the practice of bidding at bridge: "Pass," "Double," "One spade."

To appreciate the variety of alternative performatives often possible in any one particular situation is to realize that the meaning of the expression used is relevant in a more specific way than has been noted up to this point. The expression used must be one which, by virtue of its meaning, alludes not merely to the relevant practice in the context but, more specifically, to one of the alternative actions which that practice then and there allows. Thus, "I'll do it," said in a certain manner upon a certain occasion does not merely allude to the practice of promising; it alludes specifically to the act of *making* a promise. "I won't do it" can be used in exactly the same way except that it alludes to the act of refusing to promise. All this is a matter of the meaning of the expression; reasonable mastery of our language consists, in part, of being able to recognize such allusions.

It now remains for us to consider at least a few of the obvious kinds of examples which might raise questions or bring out interesting features of the analysis I have presented. My purpose here is not to provide an exhaustive defense against all *prima facie* puzzling examples. I aim instead to show the general lines of such a defense and thereby to show that the analysis I have presented cannot merely be saved but that it resolves and illuminates such apparently difficult cases.

From the standpoint of philosophical controversy over performatives, perhaps the most important distinction which needs clarification is that between the performative and what Austin called the constative. In saying something true or false about anything we are normally doing something further *in* saying what we do: we are telling or informing or reporting. Since we do not use language pointlessly, our use of expressions or sentences always has what Austin calls illocutionary force; we are doing something *in* saying what we say (not as the effect of what we say).[14] The question which perplexed Austin was whether there is any such doing which is distinctively performative rather than constative. And as far as I can see he ended by failing to find a basis for any such sharp distinction. He proposed instead a "programme" of mapping out various "families of related and overlapping speech acts." However, the analysis I have now presented does enable us to discriminate sharply between the performative use of language on the one hand and, on the other, the use of language to do something but not to do it performatively. It will be useful to state in a little detail wherein this difference lies.

Probably more often than not the task of deciding between constative and performative use is quite easy if we accept the account I have presented. This is because so much of our use of language to refer involves reference to things which are not necessary elements of the speech-act situation: "The July 21 stock averages were up a point from the day before." It is intuitively clear that this is typically not a performative; and it is equally clear that it is not performative if we apply the test of reflexive reference. But of course the writer does do something in saying it: he reports the news. It is true that at times when we give information we use an expression to refer to what happens in fact to be a necessary element of the speech-act situation though we don't refer to it *as such*; for example, "Santa Barbara has a moderate climate," said by one who is at that moment in Santa Barbara. Here again there is no problem if we accept the account of performatives as reflexively referring. There are also those cases where we do give information and in the course of which there is reflexive reference ("I", "here," etc.). But in these cases the reference, although reflexive, is not specifically to the utterance; it is not performance-referring.

What is very rare indeed is to give information and in the course of so doing to refer to the utterance itself and as such, i.e., to refer to the utterance reflexively. This is rare, but feasible: "The utterance by means of which I am now addressing you is in grammatically correct form." Notice how odd and esoteric an example one must construct in order to get something that is *clearly* not a performative and yet at first *looks* as if it might be so classified according to my criteria. However, in the light of the account I have given, even this odd remark does not really qualify as a performative, as intuitively we see it should not.

---

[14] Austin, *op. cit.*, pp. 98 ff.

It is true that in saying "The utterance by means of which I am now addressing you is in grammatically correct form," I am reflexively referring to my utterance, and it might be claimed that I am tacitly alluding to the practice of informing. But in the performative the role as marker is of the essence. That is, performance-reference is *required* by the practice according to which the utterance counts *also* as a further act. In the case of our example, the practice of informing is relevant but does not *require* for its exercise that we use expressions in a performance-referring way, though the sentence in question does *happen* to be so used. For an expression to be used performatively, performance-reference must be essential, not accidental. To put the point more informally: In the present case what I am mainly calling to your attention is the grammatical correctness of my remark and not, as in the case of apologizing, the fact that I have indeed then and there uttered certain words.

An important kind of confusion closely related to the preceding discussion is Austin's failure to distinguish between what we do when we state something and what we do when we use the expression "I state that" (and its near variants). The constative statement "He did not do it" is true or false. But Austin holds that "I state that he did not do it" is also true or false and its truth or falsity is investigated in just the same way as that of "He did not do it." Both, says Austin, make the "very same statement." And both, he adds, have illocutionary force. Yet one looks more like a constative, the other like a performative (which is nevertheless explicitly a "statement"). It appears to Austin that there is no clear distinction between constative and performative. His failure at this point to keep constative and performative distinct from one another is clearly connected with confusion in his analysis of "I state that." Let us see how this is so.

The illocutionary force of "He did not do it" may well be merely to state that he did not do it. But the illocutionary force of "I state that he did not do it" is not *merely* to state this. In addition to this, and by virtue of the performative use of the expression "I state that," its force is to establish which of several relevant practices I am exercising when I say "He did not do it." Sometimes I am putting it on the record that what I said was to be taken *only* as a statement and not, as might otherwise have been supposed, as a bet or a warning. Also it is quite common to use the expression "I state that" performatively in a situation where I am not merely stating but am also explicitly or tacitly swearing to something, or where I am invoking a prior oath to speak truly. For example, the witness in court is pressed by his interrogator:

Q. "Do you now state—remember you have sworn to tell the truth—that he did not do it?"

A. "Yes, sir. I state that he did not do it."

Thus Austin was not confused when he said that the use of "I state that" is "on a level with" such expressions as "I warn" and "I bet." But he was wrong, and crucially so for his own purposes, when in effect he also took it that "I state that he did not do it" is on a level with "He did not do it," and that therefore there is no clear dividing line between performative and constative.

Let us turn now to some other troublesome cases such as Austin's: "I warn you that the bull is about to charge." Here "I warn you . . ." suggests the performative, and ". . . the bull is about to charge" suggests the constative. For Austin this was a critical problem, but for us there is no problem. We have explicitly allowed that the same expression or sentence may be simultaneously used in different ways. It is clear, both intuitively and also by application of our criteria, that this sentence has a typical performative use (to warn). And it is noteworthy how decisively we can differentiate, from among the several simultaneous ways of using one sentence, that particular use, if any, which is performative.

Though I emphasize the decisiveness with which we now can identify what is and what is not performative use, this should not blind us to the fact that it is not uncommon to *intend* to be unclear about whether one is using language performatively or not. I may, for reasons of tact or fear, carefully speak ambiguously with respect to whether I'm warning you or merely informing you of my intention when I say, with just the right tone and manner, "I'll be there." It is of course an important part of the art of language-use and social intercourse to know how to speak obscurely or ambiguously upon occasion. However, our analysis enables us to identify such use of performatives and to understand the essential machinery of it. We are ambiguous or vague with respect to performativeness or non-performativeness when we let it be unclear whether or not we are using the expression in question in a performance-referring way, or when we let it be unclear whether or not any allusion is made to a relevant practice or to a specific act within the practice, or when we let it be unclear which practice or act it is to which we allude.

Some special remarks on one further set of examples must serve to conclude this critical discussion.

"Redouble!"[15] "Out!" In certain typical bridge or baseball uses, these are clearly performatives by our criteria as well as intuitively. At the heart of their use is the essential reference to the utterance as a marker identifying an act, an actor, a place and a time, all of which thereby also serve to identify a further act within a relevant practice. However, they are examples of a formalized performative use which calls for brief comment. In the case of "Out!" there is a deeply rooted convention that, when uttered by the umpire at the critical moment, the utterance alone will be adequate evidence that the umpire was using "out" performatively—only the most fearless (or foolish) of umpires would deny that he had "meant" it to count. In the case of "redouble" it is the rule that uttering the word at the appropriate moment counts definitively as a performative use of the expression "redouble," yet it may not have been the player's intent to use the expression in a performance-referring way. Must we here abandon the performance-referring use as central to the performative? I think not, for it is clear that it is indeed the utterance as performance of an act which remains central; we merely hold that anyone who has sufficiently mastered the game or practice to reach this particular point must be willing to take responsibility for his utterances to avoid so far as possible any ambiguity or vagueness at this stage of the proceedings. He must abide by the special rule that any use at all of the expression on certain occasions counts as performance-referring. This is akin to what the lawyers call strict liability, where the performance of a certain act can count legally as decisive evidence of wrongful intent. In both law and games, "strict liability" for one's utterances or other deeds would make no sense were it not the case that "liability" is in general *not* "strict."

Finally, it should also be noted in connection with the examples just discussed that although the practices in question happen to call for a particular formula, use of a formula is in itself never enough for performativeness. It is essential that the expression either be used in a performance-referring way or be counted according to special rule as a performance-referring use. Formulae are like other expressions in that they need not always be used in a way which is or even counts as performance-referring: the expressions "out" and "redouble" are probably more often than not used non-performatively in baseball and bridge talk.

University of California, Santa Barbara

Received January 6, 1966

---

[15] The example is from Strawson, *op. cit.*

American Philosophical Quarterly
Volume 4, Number 1, January 1967

V. PHILOSOPHICAL IMAGINATION

ARTHUR W. COLLINS

WHEN a philosopher claims that something is imaginable he usually wants to convince us that a statement asserting something contrary to what he imagines is not an analytic statement. His argument has this force: Contingent statements are those that exclude a possible state of affairs. If a state of affairs is imaginable, it is possible. Therefore, if one can imagine a state of affairs, no matter how remote from what is in fact the case, one can show that any statement that excludes what one imagines is contingent and hence is not analytic.

Let us take as an illustration an ingenious fantasy used by P. F. Strawson to show that a statement asserting the dependence of a person's experience on a certain body, namely his, is not an analytic statement. What Strawson specifically finds to be contingent is that our visual experience is dependent in three ways on one and the same body and not, and this is what he succeeds in imagining, dependent in one way each on three different bodies.

> It is a contingent fact that it is the same body [upon which a person's visual experience is dependent in three ways]. *For it is possible to imagine the following case.* There is a subject of visual experience, *S*, and there are three different bodies *A*, *B*, and *C*.[1]

With many engaging details Strawson explains how the experience of *S* would be dependent on *A*, *B*, and *C* in his imagined case.

Strawson wants to emphasize the contingency of the dependence of experience on the body because the status of this dependence is important for his concept of a person. Abstracting from Strawson's goals, let us confine our attention to his appeal to the imagination in establishing contingency and neglect the reasons for which such a conclusion might be desirable or undesirable.

Superficially I am tempted to say that what Strawson says we can imagine, we can. But is it not odd to say that I am tempted, only tempted, to concede this? For if I can say that I am tempted I suggest that there can be some doubt. How could such a doubt be resolved? How can we tell whether or not we can imagine something? How could we try to prove that Strawson cannot imagine what he says he can? These are ticklish questions. We do not ask a man whether he is sure that he has imagined something he says he has imagined, just as we do not ask whether he is sure he believes what he says he believes, or that he has dreamt what he says he has dreamt. He may lie about these things but there is no sense given to his being mistaken. There is nothing for us to assure ourselves of but his good will.

This tack will not do. We want to exclude some claims to imagination on principle. If we do not, we shall hardly be able to rest arguments on the claim that something is imaginable. We are not dealing with a sense of "imagination" and "imagine" in which the crucial difference may be that Strawson is more imaginative than we are and, aided by his more powerful imagination, he is able to see that states of affairs are contingent about which we remain uncertain. Strawson's claim is not that he succeeds in imagining where others fail. His claim is that such and such is imaginable. The sense of imagination employed here seems to be a special one. For in non-philosophical contexts we can be unable to imagine that a friend of ours should behave rudely, or unable to imagine how a certain equation would look plotted on a co-ordinate system, and we may be more imaginative than others, or less, even to the point of being unimaginative. But none of this applies in the philosophic usage.

Strawson imagines our visual experience dependent on three discrete bodies rather than on one. Can we also imagine our experience dependent on no bodies at all? If we can succeed in imagining this, we can follow the pattern: "Of course our experience is causally dependent on our having at least one body, but this is a contingent fact. It is the case that people have bodies but they might not have had. This way we would be able to refute any theory which made it an analytic proposition that corporeal characteristics can be ascribed to persons. One of the theories refuted would be Strawson's:

---

[1] *Individuals* (London, 1959), p. 90. Italics added.

> So far I have said that the concept of a person is to be understood as the concept of a type of entity such that *both* predicates ascribing states of consciousness *and* predicates ascribing corporeal characteristics, a physical situation, etc., are equally applicable to an individual of that type.[2]

Since it would follow that his own analysis of "person" is false, Strawson will not want to allow that we can imagine that people might have had no bodies at all, nothing, that is, in virtue of which we could ascribe corporeal characteristics to them. That people might have been entities with experience yet independent of any body must be unimaginable for Strawson.

All of this merely shows that there is a point in saying that we are *tempted* to agree that Strawson's story is imaginable. For our temptation, I think, would clearly extend to the imaginability of people with no bodies at all. If we want to choose between these and find three-bodied dependence imaginable and no-bodied dependence unimaginable, we must have more than temptations to go on.

I. Imagining and Picturing

Picturability appears to be the issue for many imaginings. We can imagine giants and four-armed men. We can imagine giant triangles. Can we imagine a four-sided triangle? Well, there are no giants but we know just what a giant would look like, but it is not merely that there are no four-sided triangles. We cannot even draw a picture of one. This analogy is weak because a picture of a four-sided triangle would *be* a four-sided triangle, so that if we could even form a picture such things would exist in fact and not just in imagination.

This partial failure of the analogy does not matter. For in spite of some intuitive agreeableness the explication of imagination in terms of picturability fails completely. To begin with, for most imaginings a picture is gratuitous. And in many, if I form a picture at all, I need not form a picture of what I imagine but of something else. "Imagine that you are a great scientist." Here I may picture myself surrounded by apparatus, or receiving the Nobel Prize. I can say that I have imagined that I am a great scientist even though I deny that I have pictured myself in any way at all, or even if I admit that I pictured myself just as I would have if asked to imagine myself a mediocre scientist (surrounded by apparatus) or a great novelist (receiving the Nobel Prize). Furthermore, the relevance of those instances of imagining in which we do feel that pictures ought to be involved is not at all obvious. We know from many useful pictures that a picture does not have to be much like what it pictures. Thus, Figure (a) is a picture of a three-dimensional space, Figure (b) is a picture of two lines with a common point and both parallel to a given line, and (c) is a picture of two parallel lines meeting at infinity. We must concede that our intuitions concerning pictures break down when we consider these pictures, and these pictures already have a use.

Whether something can be pictured or not depends on our willingness to use some picture as a picture of that thing. We might, for example, find (d) offered as a picture of a four-sided triangle. In (d) the dotted line is a device to help us digest the picture without being too upset by it. It functions just like the dotted lines in (a) through (c).

Fig. (a)

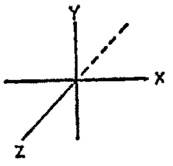


Fig. (b)

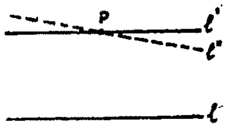


Fig. (c)

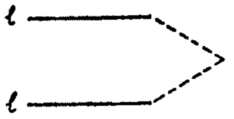
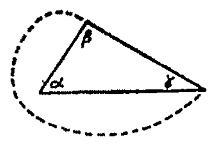


Fig. (d)



I do not argue that four-sided triangles are imaginable after all. The picture (d) is interesting only because it is no harder to construct than the useful pictures (a) through (c). Obviously, if we want a further criterion, it will be one with the help of which we can dictate that (d) is not a picture of a four-sided triangle. We shall not be able to draw this conclusion from the picture itself. Picture (d) does not picture a four-sided triangle because there could be nothing fitting that description for (d) or anything else to picture. Now it is not our ability to form pictures that is in question but the propriety of using them in certain ways. We cannot use a picture as a picture of a four-sided triangle because the very concept of such an object is contradictory.

---

[2] *Ibid.*, pp. 104–105.

II. Imagination and Contradiction

Shall we say, then, that it is the need to avoid contradiction that limits our imagination? One thing is not accounted for by this suggestion. If the criterion for imaginability is freedom from contradiction, why should philosophers argue that something is contingent because it is imaginable rather than because it is not contradictory? The only good reason that can be given for this order of philosophical argument is also a good reason for not regarding freedom from contradiction as the criterion for imaginability. Plainly, imaginability itself is being employed in contexts like Strawson's story in order to demonstrate freedom from contradiction. What is imagined serves as a model for the simultaneous truth of a number of propositions just as a model suffices for the proof of consistency of a set of postulates. The imaginary model is offered as a rough proof of the consistency of the assertions made within the story that we find imaginable. What we imagine *could* be true and hence could not be contradictory.

That the model exists in fact is crucial for model proofs of consistency. The philosopher's model, however, is only imaginary. It seems a sensible caution to wonder whether we are not merely imagining that the story that we tell could be true, merely imagining that this state of affairs is possible and, therefore, merely imagining that it involves no contradictions.

We may agree that one cannot imagine an object with explicitly contradictory properties or a state of affairs described by explicitly contradictory sentences. But it is not clear that one cannot imagine that some object or state of affairs that in fact does involve a contradiction does not involve a contradiction, or imagine states of affairs that involve a contradiction *implicitly*. Against an effort designed to show that we cannot imagine what we claim that we can, we can simply refuse to let our attention be directed to the contradictions which can be shown by a more or less lengthy argument to be involved. Everyone is familiar with the fantastic idea of time travel by time machine. Shall we accept time-travel fantasies as imaginable and argue thus: "In fact we are not able to go back in time. But this is only a contingent truth since I can imagine a time machine in which I could travel both to the past and to the future?"

The nature of the problem is now becoming clearer. For my part I think it obvious that I can imagine a time machine but I am not in the least inclined to suppose that such a machine is in any sense possible. Do I mean that it is not even logically possible? Do I mean that my imaginary time machine is contradictory?

Arguments may be designed to try to overthrow my claim that I can imagine time travel. What do I imagine exactly? Perhaps a sort of windowless flying saucer. I can enter it, set the controls, and—I am in the fifth century B.C. Now is there not a contradiction in this story? I enter the time machine. Then, later in time, I set the controls; later still in time, it is the fifth century B.C. which is earlier in time. Therefore, my story involves a time which is both later than and earlier than another time and, consequently, I cannot really imagine my time machine at all.

In my story I do not claim that I enter my time machine and then later in time visit the past which is earlier in time. I enter the time machine and then, later in the story and later in time, I set controls and then, later in the story but not later in time, it is earlier in time and I am there. Naturally I tell about this, first telling the bit about my entering the time machine and then telling the bit about my being in the past, but I do not have to say that I was in the past later than when I was in the present. At least this makes much more work for the sceptical. But is it only because, for the moment, I can stave off this effort to elicit a contradiction from my time-travel story that I can claim to imagine time-travel at all? It is here that conflict arises with an understandable appeal to my right to say what I can imagine and what I cannot. I have imagined my time voyage. Further argument about that should be pointless.

I agree that I cannot imagine what is contradictory but at the same time I insist that I can imagine *this* even though I cannot say that I know that it is not contradictory. To shed light on the obvious conflict of criteria here we need one more illustration. Let someone imagine that the square of 14 is 197. This example is chosen to make the search for a contradiction easier. That way we can see what happens when we succeed in finding one. We try to overthrow the claim about the square of 14 by asking just what is supposed to be imagined:

A: One day someone discovers by computation that 14 times 14 comes to 197. Others check this.

Then word gets around and for a while everyone is amazed that they all seem to have had it wrong before.

Q: But do you imagine that other arithmetical relations remain as they are and that the procedure for computing the squares of numbers is retained? You know that if you don't imagine it this way you will merely have changed the ordinary meaning given to these things and not have imagined what you claim at all.

A: I do imagine other arithmetical facts and the procedure for computing squares to remain as they are.

Q: Then do you admit that what you say you imagine is contradictory? For the computation by which you imagine it to be discovered that the square of 14 is 197 begins, as it does now, four times four, and this product is not imagined to be different and hence is 16 and the last digit of this product will be the last digit of the square of 14 because the method of computing squares remains the same. Therefore, the last digit of the newly discovered square of 14 will be 6 but the last digit of 197 is not 6. In the story that you find imaginable the last digit of the square of 14 both is 6 and is not 6.

A: Imagining a number which ends in 6 and is also 197 is very hard. Perhaps it is impossible. But that is beside the point here, since I am only trying to defend the much easier fantasy that 14 squared is 197. You ask me whether 4 times 4 still equals 16. This is irrelevant but out of good nature I extend my imagination a bit and imagine that it is. You ask me whether the new computation begins by forming this product. Irrelevant again, but let us imagine that it does. We can imagine whatever we like about these things. The important thing to concentrate on is not imagining the beginning of the computation nor imagining the process of computation; we need only imagine the outcome. And that I imagine this without very much trouble to be 197.

But you are unhappy with this. You find yourself capable of imagining a time machine in which one can visit the past and the future. Suppose that interrogating you on your claim I ask, "Do you have to drive the time machine as it goes through time or is it automatic? Do you have to set the time destination in advance or can you change it en route? Do you start the time machine? If the fuel tank is half full when you start for the past, is it full or empty when you get there? Can you alter the past when you get to it? If not, how do you know that it is not merely a documentary reproduction? If so, could you change it so that a time machine could not be invented, at least not in your time?" All of these questions will be designed to catch you in little inconsistencies. They will aim at doing this by obliging you to imagine more and more details which will then present the problem of agreeing with one another.

Now will you let yourself be put in the position of a criminal shoring up his alibi, forced into fresh lies by the persistent questioning of the police? No. You will not be put off by all of these questions and all of these additional demands made on your imagination. You will rightly insist that you do not claim to imagine the construction of a time machine, nor its mode of operation, nor what one can and cannot do as a consequence of travelling in time. You claim to imagine only the use of the time machine. Similarly, I do not claim to imagine the steps in the computation that lead from the problem, *What is the square of 14?* to the solution 197. Nor do I imagine, or claim to imagine, whether or which other mathematical or non-mathematical facts are different in the world in which 14 squared is 197, although I could do some of these and might not be able to do others. In the realm of imagination our stories are intentionally incomplete so as not to be fettered by contradictions that might arise were we to try to complete them.

In one sense it is easier to imagine a time-voyage than to imagine that the square of 14 is 197. This is in part because we are accustomed to having machines do astonishing things for us without our being astonished and without inquiring how the machines get those things done. Our acquaintance with elementary arithmetical operations, in contrast, is not greater than our ability to perform them ourselves. In another sense, it is the time-travel fantasy that seems the less digestible. For we can understand what it is that the imaginary computation is supposed to achieve: it is to reach the answer "197." But the difficulty for time-travel stories can be focused entirely on the outcome. We could imagine that a time machine could *do it*, if we could imagine what it is that the time machine is supposed to do. The suggestion, "It is supposed to go to the past," is as taxing as, "It is to reach the answer '197'," is not.[3] In both cases, however, the contention that what is imagined is possible would be rebutted by rooting out inconsistencies which the imagination, in order to succeed, has ignored.

---

[3] For this distinction I am indebted to a comment by a referee for the *American Philosophical Quarterly*.

Imagination is a useless tool when it comes to proving freedom from contradiction. Intelligent men sometimes assert and assent to propositions or sets of propositions which contain logical incompatibilities. Contradictions are then *believed*, not merely imagined. We must put up with an imperfect capacity to recognize them. When they are only implicit, contradictions even in mathematics are not always an impediment to belief. This is why consistency proofs in mathematics have a point.

III. Vagueness and Contradiction

Suppose I tell a fantastic story using a descriptive term *F*. Suppose you know, because it is an obvious inference from the usage with which you are familiar and with which you rightly think me to conform, that it is part of the meaning of *F* that *F* only applies to things to which another term *G* applies. In other words, if *G* is not true of an individual, then, by definition of *F*, *F* is not true of that individual either. But further suppose that it is not an obvious implication of ordinary usage or of my usage that applicability of *G* is, in virtue of the meaning of *F*, a sufficient as well as a necessary condition for the applicability of *F*. Nor is it an obvious implication of that usage that *G* is not a sufficient condition. Finally, Let *H* be a third descriptive term such that *H* is true of all the real individuals of which *F* is true.

To bring this schema to bear on contradiction and imagination, suppose that a logical consequence of the fantastic story that I tell, using the concept *F* among others, is an assertion ascribing *F* and denying *H* to one and the same individual. Someone who wants to prevent me from putting the "possibilities" revealed by my imagination to work will seek to show that my story is contradictory by showing that the correct analysis of *F* suffices to establish the falsehood of my ascription of *F* and denial of *H* to the same thing. He may use his imagination, too. For example, "Imagine that you took a pill and that as a consequence of taking it you could read Latin." This may be supposed to be part of an effort to analyze the concept "learning." Would we say that you had *learned* Latin? I suppose not. How does the fantasy help? Well, we want to know the features that a process ought to have in order that it properly be called a learning process; or what sorts of things it is implied that a creature has done when he is said to have learned something. Suppose every process that results in the possession of a certain *kind* of capacity, ability, or skill is in fact a learning process. Then it is not surprising if we come to think that culmination in that kind of capacity, ability, or skill suffices as a reason for calling some process a learning process. By hypothesis, no process can be pointed to as a counter instance.

Now let *F*, from the schema above, be the expression "learning" or "is a learning process" and *G* be "is a process culminating in the possession of a certain kind of capacity, ability, or skill." I do not think that I can suggest a widely acceptable candidate for *H* so I shall only indicate the area in which candidates can be found. If a man can speak a language as a consequence of taking a pill, should we say he has learned that language? The motivation for a negative reply is, at least in part, similarity with other cases where it is fairly clear that we would not speak of learning. For example, bees can build hexagonal honeycombs. Some process in the life history of bees, no doubt, culminates in this ability. But it does not seem at all right to conclude that this process must be a learning process and that the bee must learn how to do it at some time prior to exercizing the ability to build hexagonal honeycombs. Substitutes for *H* will be found in the course of an effort to discriminate between a hypothetical process which culminates in this ability of bees and certain processes in which, say, rats or people participate and which we definitely want to call learning processes. If we succeed in picking out the right *H* we will say that we have clarified the concept *F*, or analyzed *F*, here, "learning"; it now being clear that *G* and *H* are both necessary conditions for the applicability of *F*, in virtue of the meaning of *F* alone.

Suppose that someone produces such an analysis of *F* in order to counter my fantastic story: "Because *H* is not applicable, we do not say that the accomplishments of bees are learned, and because *H* would not be applicable, we would not say that a man who had only taken a pill had learned Latin. In short, it is part of the meaning of *F* that if *H* fails to describe something, so does *F*." Then my imagined possibility is contradictory, since it entails that *F* describes something that *H* does not. The concealment of the contradiction has hinged on the vagueness of *F*, and the difficulties encountered in bringing the contradiction to light are at least as great as those involved in explicating *F*.

We talk as though the problem can only be difficult, not insoluble; as though F has either got to have or not have a certain component of meaning; as though we await only the correct analysis of F to decide which. But such analyses cannot always be expected for the important vocabulary of fantastic imaginings as well as important vocabulary in other everyday and philosophical contexts. Whether H is always true of whatever F is true, by definition, or only in fact, cannot always be decided by analyzing some domain of usage of F *unless one is willing to include contexts like the telling of imaginary stories as part of that domain.* It is in order to facilitate such decisions that philosophers appeal to fantastic imaginings. An imaginary story minimally pretends that it is not contradictory and hence that any analyses that would make it contradictory are not correct, and that any analyses required for its consistency are correct. Plainly, we cannot credit these pretensions if our analyses are to decide the consistency of these stories in any interesting way. Therefore, the usage which we analyze must be confined to the domain of reality. But in the domain of reality the vagueness of concepts is often, so to speak, part of their meaning.

Let us return to my term F. I use F and you want to know whether I am using it so that G alone must apply to justify the application of F, or so that G and H must both apply to justify the application of F. You ask me and I cannot or will not answer. Then my usage is vague because there are at least in principle circumstances, namely those in which G applies and H does not, wherein we do not know whether F applies or does not apply, although we have all the relevant facts. How shall I dispose of the charge of vagueness? Must I make an arbitrary choice, a stipulation?

I say, "As a child I learned Latin." You want to know whether my concept of learning is such that I would say the same thing if what I had done as a child had been to take a pill, provided that the outcome had been the same capacity to perform. Your question presupposes that when all the facts are in I ought to be able to answer. There is nothing else to find out and, after all, it is my concept. Or speaking of ordinary language, of concepts in the public domain, a philosopher should be able to answer such questions. They should be answered by philosophical analyses.

This presupposition is not very convincing. I want to do it justice. I cannot evade the question on the ground that there are no such pills. If there were I would have to decide. Therefore, I can be asked *now* what I would say *then*. If nothing presses me to make one choice rather than the other I can choose arbitrarily. But I have some grounds. The analogy with the abilities of bees suggests that I withhold "learning" from acquisition of languages by taking pills. And another pill might give me the capacity to digest cellulose. Since that would not be learning, I decide that taking a Latin pill would not be learning Latin either. On the other hand, the Latin pill might be only the final triumph of a long pharmaceutical development. First there are pills for students that make it just a shade easier to learn Latin. . . . So had I better allow that the pill taking would be learning? Imagination can give us reasons for either choice, so the choice remains arbitrary.

There is no point in making an arbitrary choice and no pressure to make one just because we have a theoretical option. If all of the situations in which G applies are also situations in which H applies, we can simply ignore the other, imaginary situations that may, or may not, in principle be possible in which G applies and H does not. In general, fantastic stories are neither endorsed nor ruled out by the meanings of terms as ordinarily used. Analyses which would decide the consistency or inconsistency of a fantasy do not lie concealed in our language as we use it and need not be made artificially by arbitrary choice. Interesting illustrations of choices that determine meanings more or less conventionally can be found, for example, in law or the history of science. But they do not illustrate a brute impulse to clarify. They are forced choices pressed on us by the need to decide between contending explanations, or contending theories, or by new discoveries, or by needs for practical decisions which render inadequate the system of distinctions marked by prevailing usage.

Imaginary discoveries and fantastic alternative explanations do not force us to make decisions about meaning. In elaborating an imaginary story, we are encouraged by an intuition that the question of consistency *can* be decided arbitrarily in some way so that our story will not be contradictory. An analysis excluding H from the definition of F can be adopted arbitrarily. Then the consequence of my fantasy would only be false, not contradictory. Furthermore, it is natural to think that if a contradiction cannot be derived from statements in virtue of the meaning that they now have, then those statements are consistent. Will many defend this view explicitly? It might be argued that if we can make a set of statements explicitly consistent by adopting definitions which do not run counter to existing usage but merely delineate, extend, and perhaps slightly reform that usage, then part of what we are making explicit is the consistency that those statements, however vague, already possess.

If the motivation for definitions, however, is only the preservation of consistency of an imaginary story, then adopting them is not legitimate. The story does not need to be thus rescued to be imaginable, and support accrues to a spurious claim that the analyses involved have some philosophic interest.

If we insist on retaining a principle, "Every statement and every set of statements is either consistent or not consistent," then we have also to suppose that we can always say in advance how a term would be used in the face of unexpected future or fantastic contingencies. But what counts as a contingency, that is, as a describable situation with respect to which we can be asked how we would use some expression, depends on the analyses of the concepts employed. We shall not feel obliged to explain how we would extend our usage to handle contradictory states of affairs. And whether or not a description of a state of affairs is contradictory depends on how we would extend our usage.

Since we cannot in general provide the analyses that would be decisive, we cannot in general say that a fantasy is contradictory or not contradictory. And since we cannot always pronounce on the analyticity of statements which exclude that which is asserted by the fantasy, these statements will be neither analytic nor not analytic.

I think it is not insignificant that Mr. P. F. Strawson, whose use of imaginability to establish non-analyticity we have cited earlier, takes a tolerant view of the difficulties that logicians have run against in their efforts to make explicit a distinction between analytic statements and those that are not analytic. Specifically, Strawson, together with P. G. Grice,[4] argues in a context of opposition to Quine that our failure, even after many very strenuous efforts, to provide an explication of the distinction between analytic and synthetic statements does not suggest that there is no distinction to be drawn. That there is a distinction is indicated, according to Strawson and Grice, by the fact that different philosophers will independently classify statements in roughly the same way with regard to analyticity. This agreement is only to be explained by supposing that the philosopher's are detecting some distinction, however difficult they or other philosophers find it to say clearly what that distinction is.

I do not contend against this argument for the existence of a non-explicated distinction and against Strawson's tolerance. It is very important to emphasize, however, that philosophers' agreement in classifying statements as analytic or synthetic must be supposed to be rough. There will be three classes of statements produced: Statements that all or most philosophers call analytic, statements that all or most philosophers call synthetic, and statements on which philosophers disagree rather widely. What kinds of statements can we expect to find in the last class?

We shall find that the statements which are of interest to philosophers, statements which express philosophical positions, which contain wholly or in part philosophical analyses, which assign a status to ethical principles or principles of science or to mathematical statements, are virtually all statements that will appear in that perhaps narrow area in which philosophers when classifying statements as analytic or synthetic do not agree. Is this too obvious a claim to defend? A glance at the recent history of philosophy reminds us that philosophers have been unable to agree on the classification in point of analyticity of philosophically interesting statements like, "$7 + 5 = 12$," "The angle-sum of a triangle is equal to 180°," "Force equals mass times acceleration," "Insofar as something is productive of pleasure, it is good," "Thought requires brain processes," "Every event has a cause," "You cannot go back in time," and "My experience is causally dependent on this particular body." Disagreement has been magnified by new ways that are devised for reinterpreting interesting statements, e.g., natural laws as rules, so as not to have to classify them at all, and some philosophers will just plainly refuse to classify some of the statements.

It is chiefly because we all know on what list, "All bachelors are unmarried," will appear, and on what, "The cat is on the mat," that we can agree with Strawson and Grice that there must be some distinction to which philosophers succeed in conforming approximately. But supposing we do accept that there is a distinction, we cannot use the distinction in connection with interesting statements unless we have more than this classifying agreement. On just these interesting statements we do not find agreement.

---

[4] See Strawson and Grice, "In Defense of a Dogma," *The Philosophical Review*, vol. 65 (1956), pp. 141–158; the dogma: one attacked by Quine, "Two Dogmas of Empiricism," in *From a Logical Point of View* (Cambridge, Harvard University Press, 1953).

Philosophical preoccupation with the out-of-the-ordinary, preoccupation that leads us to take seriously fantastic "possibilities," has the same source as philosophical concern with analytic truth. By the special philosophical concern with analytic truth I mean the concern to show that sentences like those quoted above and most others that are philosophically interesting are or are not analytic, and to use the notion of analytic truth as a tool for refuting and establishing philosophic positions. Analyticity promises a nearly magical power to guarantee philosophical generalizations. Crudely, if philosophers analyze concepts and if their analyses are good ones, then their positions will be analytic truths. The promise of this point of view has its source in the same power that is supposed to enable us, by use of the imagination, to decide questions of possibility, consistency, and meaning. For both it is required that we think that language as we already use it is committed in ways in which it is in fact not committed. We can only show that certain fantasies are consistent or inconsistent by making choices among alternative meanings that *could be assigned* to some of the terms involved. But within the context the only motive for making the choices is the motive to make the fantasies consistent or inconsistent. Similarly, we can often only defend the judgment that a philosophically interesting statement is or is not analytic by making decisions that are only motivated by the urge to defend that judgment. These motives are not good enough.

Columbia University

Received January 20, 1966

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 1, January 1967

VI. THE SEMANTICS OF IMPERATIVES

ERNEST SOSA

CAN there be a logic of imperatives? The logical relation between "He grins and bears it" and "He grins" is universally recognized to be one of logical implication. Is the relation between "Grin and bear it!" and "Grin!" the same in also being one of logical implication? Whether indicatives and imperatives are indeed the same in this respect or not, let us assume that the preanalytic similarity of such cases as the two just considered does give at least *prima facie* reason for believing that they are both cases of logical implication. Before we can have *conclusive* reason for so believing, however, we must surely develop some account of what it is for the members of a set of imperatives logically to imply a further imperative.

We say that the members of a set of indicatives logically imply a further indicative provided that the members of the set cannot all be true while the further indicative is false. Clearly, this will not do for imperatives, since no imperative can ever be true or false. Our main problem in what follows is: What *will* do for imperatives?[1]

I. THE USES OF IMPERATIVES

Imperatives have many different uses: to command, to instruct, to request, to plead, to appeal to a rule or law, and many others that anyone with a thesaurus can look up. And obviously these do not all have the same "logic," in a broad sense of the term.

In spite of the marked differences in the various uses of imperatives, however, I should like to propose a distinction between declaratives and directives, and to develop a logic of directives. It ought not to be forgotten that declaratives too have several different uses: to assert, to describe, to outline, to narrate, to explain, and so forth, not all of which have the same "logic"; and that in the logic of declaratives we nevertheless abstract from all of these various uses so as to study logical features shared by all declaratives, so long as they are used "assertorically" (which encompasses a broad range of uses including those just mentioned).

It is to emphasize our concern with a functional, and not merely syntactical category that I have chosen to speak of "directives" rather than of "imperatives." Let it be clear, however, that directives are simply imperatives whose uses are restricted to the normal ones of commanding, instructing, requesting, and so forth; and whose uses exclude, for instance, that of emphatically asserting that water boils when heated through an utterance of "If you want water to boil, then heat it!"

It may well be that we are here doing logic at a remove: that just as standard logic deals primarily with propositions or statements and only secondarily with indicative sentences, so the logic of the practical deals primarily with norms or prescriptions and only secondarily with imperative sentences. Nevertheless, it is convenient to focus on the sentences, once having made it explicit that we are here concerned with their normal or typical uses.[2]

II. THE LOGICAL STRUCTURE OF DIRECTIVES

We may always translate any use of a directive so as to reflect the distinction between its propositional core and its purely directive "modality." My command, or suggestion, or request that you mind your own business, for example, can be conceptually decomposed into the propositional core: that you mind your own business, and the directive modality of my suggesting or my commanding or my requesting. There appears to be no provision in ordinary language for any determinable under which these all may fall as determinate modalities, and so I have chosen the symbol "*D*: " for such a purpose, where "*D*: " is to be read "Let it be the case: ." Accordingly, when considered as "for use as prescriptions," imperatives will in this study take the form "$D{:}\alpha$", where α is the propositional core of the imperative. It should be provided, however, that each of the propositional variables in α can be replaced only by nominal clauses describing possible states of affairs which are exhausted by someone's making something be true.

---

[1] I gratefully acknowledge receipt of a grant for work on imperatives from the Canada Council. I am also indebted to Professors Nuel Belnap, Nicholas Rescher, and John Robison for their helpful comments.

[2] Cf. *Strawson* (1964), pp. 31–32. For references of this form, see the Bibliography appended at the end of this paper.

This general view of directives is nowadays associated especially with the name of R. M. Hare and the terms *neustic* and *phrastic*. It can be traced, however, as far back as Frege,[3] and even Descartes.

It is widely known that Frege believed uses of indicative sentences to comprise both expression of a content and assertion of this content. What is less well known is that, as Rulon Wells has pointed out, in his later writings Frege went on "to identify the sense of a question with the sense of its indicative counterpart. . . . Presumably, then . . . interrogative and imperative sentences differ from indicatives only in virtually containing an interrogation sign or an imperation-sign in place of the assertion sign."[4]

"Of my thoughts," says Descartes,

> . . . some are, so to speak images of the things and to these alone is the title "idea" properly applied; examples are my thought of a man or of a chimera, of heaven, of an angel, or [even] of God. But other thoughts possess other forms as well. For example, in willing, fearing, approving, denying, though I always perceive something as the subject of the action of my mind, yet by this action I always add something else to the idea which I have of that thing; and of the thoughts of this kind some are called volitions or affections, and others judgments.[5]

Now some of these entities which are subjects of emotion, and so on, will not be propositional; but some surely will be propositional. (Else what is the subject of Descartes' "denying," for instance?) In such cases, since imperatives are among the natural linguistic media through which to express "volitions," and since the volitions turn through analysis into compounds of a propositional entity which is the "subject of the action of the mind" and something else which is added thereto, we may therefore cast our imperatives into a form comprising both a linguistic component corresponding to the proposition and a linguistic component corresponding to the "added things" to which Descartes refers.[6]

It is noteworthy, finally, that the English counterparts of our practical modal sentences (those falling under the determinable $D{:}p$) could all be interpreted in one of two different ways: as declarative reports or as *sui generis* practical expressions. It is the latter interpretation that I wish to put on our sentences and, consequently, on their English counterparts. Thus, although "I command p" can naturally be interpreted as an account of certain of my activities and thus as a statement or assertion, this is not the way it is here to be understood. Instead, it is here to be taken as fully practical, as having its use in the actual giving of commands.

III. AN INFORMAL EXPLICATION OF VALIDITY

The basic datum upon which our subsequent theoretical constructions will stand is the intuitive plausibility of such arguments as:

(a) Pick up the books and the pencils.

So: Pick up the books.

(b) If prices fall tomorrow, buy (tomorrow) twice as many shares as you normally buy.
But prices will fall tomorrow.

So: Buy (tomorrow) twice as many shares as you normally buy.[7]

The usual interpretation of these arguments has been in terms of statements and orders or requests. And the semantical values for statements are the familiar ones, true and false. But what corresponding semantical values should one then assign to directives so as to make sense of such plausible arguments as those above?

---

[3] See *Hare* (1961). Hare also follows Frege with his proposals for the construction of a calculus of gerundive clauses. It is interesting that this program has actually been carried out by the logician Haskell B. Curry. See *Curry* (1964), especially the section entitled "The Propositional Interpretation," on pp. 168–172.

[4] *Wells* (1950–1951), p. 557. Wells refers us to *Frege* (1892), p. 39c; *Frege* (1893), pp. Xa and 9; and *Frege* (1919), p. 145a.

[5] E. Haldane and G. Ross (eds.), *The Philosophical Works of Descartes*, vol. I (Cambridge, 1931), p. 159.

[6] Compare *Leonard* (1959), where this kind of analysis is extended to questions, as in Frege.

[7] That directives can be "contradictory," moreover, seems not to be alien to the layman's intuitions. As an example I quote from an article, "The Foreign Aid Fiasco," by Paul Drake, which appeared in *The Reporter* (Jan. 16, 1964): "Most upsetting of all to aid supporters was the critique's careless and defensive style, the generalities and *contradictions*. . . . At one point the report spoke out against aiding governments in socialistic projects while at another it called for continuation of assistance to India, the biggest example of a country receiving assistance for state-sponsored endeavors. The administration was exhorted to tighten the terms of its loan and grants, yet to channel increased amounts through international lending agencies—where U.S. control would all but cease." [My emphasis.]

Three main proposals have been made. Some have argued that the semantical values should be "satisfied" and "non-satisfied."[8] Others have suggested instead "justified" and "non-justified."[9] And still others have proposed "in-force" and "not-in-force."[10] There are serious difficulties, however, with each of these sets of values.

What is it for a command to be "satisfied"? As a first approximation we may say that a command is "satisfied" when what it commands takes place or occurs. But this leads directly to the question of what it is that a command commands (a directive directs). And an answer to that question seems to be, again roughly, that what a command commands is that one do something on a certain occasion or on certain occasions. Suppose we take as our paradigm "that one do A at (juncture) j."[11] Roughly, then, a command commands one to *make it true* that one does A at j. But it seems desirable to distinguish explicitly between:

(c) "Making it true that-one-does-A-at-j"

and,

(d) "Making-it-true-that-one-does-A, at j."

Two different conceptions of a proposition are involved in (c) and (d). (c) presupposes the spatio-temporally "closed," and (d) the spatio-temporally "open" notion of a proposition. In what follows we shall be using (c) rather than (d). And if we take the step from (c) to the equivalent:

(e) "Making that-one-does-A-at-j be true"

we may compare this with ordinary cases of

"Making x be ϕ," where x is a logical object and ϕ is a possible attribute of that object

such as:

"Making the chair (be) secure"

or

"Making the law (be) equitable"

The following is presupposed in such cases. If one is to make x be ϕ at j then it is necessarily the case (by the meaning of the words involved) that: Given the set of conditions holding at the time, x would not be ϕ at j by itself (in the absence of one's efforts), i.e., that there exist no forces which *even as one acts* are operative and which would make x be ϕ at j, even if one did not act. It is *not* required, however, that there be no forces which, were one not to act, would then become operative; for even if there are such forces one still can make x be ϕ at j, though regardless of one's efforts x *would* be ϕ at j. It may be of help to give examples of the two kinds of cases I have in mind. The last kind of case is exemplified by my closing the door when if I had not done so John would have, since if he had not seen me stand up and approach the door, etc., then he would have done so himself. This case is to be distinguished from the one where if the efforts I exerted in the attempt to close the door had not been exerted, the door would still have been closed, *not* because if John had not seen me stand up, etc., he would have done so himself, but rather because he, John, actually had his hand on the door just as I did, and went through the motions and exerted his efforts even as I did, efforts which would have sufficed for the closing of the door even in the absence of my own. Now, surely in the former case I do bring it about, by myself, that the door is closed, but not in the latter case.

The following definition has emerged from the foregoing.

> One makes p be true if and only if one's efforts are *ceteris paribus* necessary and sufficient for p being true; and p is true; and there is no other agent whose efforts are *ceteris paribus* necessary or sufficient for p being true (for in that case we would have an instance of collecting making-true).[12]

It may well be asked, "Why bother to analyze the notion of someone's making something be true?" And, it may then be added: "That is not our problem. Our problem is to explicate the concept of the validity of imperative inference." It must surely be granted, however, that our explication of validity would be of little value if the notions it presupposed were as obscure as that of validity itself. Hence it seems appropriate to render such notions as clear as we can. My concern to analyze making-true may therefore be justified by the fact that it is a concern to render clear one such notion. In terms of the notion of making-true a provisional definition can now be offered for a fundamental semantical concept of the logic of directives, one which will later be used in the explanation of validity. I mean the concept of *directive satisfaction.*

---

[8] See, e.g., *Hofstadter and McKinsey* (1939), pp. 446–457; and compare the excellent discussion of deontic logic and the logic of imperatives in *Lemmon* (1965).

[9] See *Castañeda* (1960).

[10] See, e.g., *Lemmon, op. cit.*

[11] Where a "juncture" is a spatio-temporal specification.

[12] On this topic see also *Black* (1958); *von Wright* (1963); and *Sosa* (1965). Let us take this as a working definition in this paper. Actually the concept seems to be a bit more complicated than is here allowed for. Cf. *Sosa* (1965).

> A directive is satisfied provided its propositional core is true.

(It will be recalled, from Sect. II, that the "propositional core" of a directive "is exhausted by someone's making something be true.")

And now it seems very plausible to say that an argument is valid provided that if its imperative premisses are satisfied and its indicative premisses are true, then its conclusion is satisfied if imperative and true if indicative.

The intuitive and *prima facie* plausibility of satisfaction values is threatened, however, by such argument-schemata as:

(f) $\dfrac{\mathrm{D}:p}{\therefore p}$

(g) $\dfrac{p}{\therefore \mathrm{D}:p}$

In the first case, if the premiss were satisfied the conclusion would certainly be true. In the second, if the premiss were true the conclusion would certainly be satisfied. It is obvious, however, that we should not adopt any criterion upon which such schemata turn out to be valid. And so, while it may be that we can still use satisfaction values in devising a criterion of validity, we are well advised to handle them with care.

We come now to the second set of semantical values for imperatives which has been proposed in the literature on the subject. Professor H. N. Castañeda introduces the values "justified" and "non-justified," which he defines as follows:

> Let us fix a context C, i.e., a set of ends E which are pursued by both the speaker and agent X together with their logical and causal connections to the circumstances of the speaker and X. Now we can define:
>
> (D_1) The imperative "X, do A" is *justified in* C if and only if (i) Both X and the speaker subscribe to E, (ii) X's doing of A is such that at least it does not further the failure of E, and (iii) A conforms to the conventions and decisions to which both X and the agent subscribe in their pursuance of E.
>
> (D_2) The imperative "X, do A" is *non-justified in* C if and only if it is not justified in C.[13]
>
> An imperative or resolutive is *justified* (simpliciter) if the context C is made up by the highest ends the speaker and the agent subscribe to.[14]

The first thing to notice about Castañeda's theory is that no semantical values have been defined for *conditional* imperatives. But even with respect to categorical imperatives there are serious difficulties. For when X addresses Y, with the imperative "Y, do A," say, the ends that count are the ends of X and Y, none other; but when X addresses Y and Z with the imperative "Everyone do A," then the ends that count are those shared by X, Y, *and* Z. Call the first set of ends E and the second E'. Now from the fact that the imperative "Everyone (Y and Z) do A" is justified (with respect to E') it does not follow that the imperative "Y, do A" is justified (with respect to E). And so the inference from "Everyone do A!" to "X, do A!" fails on Castañeda's theory.

Of course these criticisms do not establish that there is no hope for justification values. Indeed by using these values Castañeda is able to interrelate imperatives with ought assertions in a most interesting and promising way.[15] However, until his theory overcomes these and other difficulties it must have the status of a program. And until such a program is fulfilled, and even afterward, it may be illuminating to consider alternative possibilities. We turn now to one of these.

Professor E. J. Lemmon has suggested that a proper semantical value might be "in force."[16] Directives could then be said to be "in force" or "not-in-force" with respect to the director in question. And then the criterion of validity might read: An argument is valid if and only if given that all the directive premisses are in force and all the declarative premisses are true then necessarily the conclusion is in force if it is a directive and true if it is a declarative. One of the main points in favor of this proposal is that it seems to accord with our intuitive rejection of such argument schemata as (f) and (g). For it does seem implausible that just because a directive is in force it follows that what is directed is true; or, that from whatever is true (given that it has reference to action, and so on), we can deduce that some directives are in force. By digging a bit deeper, nevertheless, puzzles are unearthed.

---

[13] *Castañeda* (1960), p. 171.

[14] *Ibid.*, p. 173.

[15] See *Castañeda* (1960 and 1962). It may be replied that justification values have not even been defined for cases of more than one agent, but this in itself would be an admission of inadequacy, inasmuch as the theory is meant to include a "quantificational" logic of imperatives.

[16] *Lemmon* (1965), *passim.*

What does it mean to say that an order is "in force"? Suppose a man gives an order and then falls dead. Does his order continue to be in force until the stretch of time ends during which it was to be obeyed? Suppose a man is not now and never has been aware of a certain imperative. Can the order of his which would be formulated by means of that imperative still now be in force? If it cannot, then what can be said about the following:

(h) Whenever it rains, close the door!
(i) It rains on Christmas Day, 1970.

(j) Close the door on Christmas Day, 1970!

Suppose the commander is not now and never has been aware of the conclusion command. Then it can't now be in force, in accordance with the assumption we have just made. But the conditional-command premiss *can* all the same be in force and the propositional one be true.

Say, on the other hand, that we do allow orders to be "in force" with respect to a certain commander even though he has never even thought of them. Then the question again arises: What is it for an order to be in force? A partial answer to this question is provided by Professor G. H. von Wright. "That an authority has prohibited something," he writes,

> . . . entails that he can and is prepared to see to it that this thing is not done. He threatens prospective trespassers with punishment, and takes steps to punish those who in fact disobey.[17]

We may then say that a prohibition is "in force" only if the authority is prepared to see to it that the thing prohibited is not done. Von Wright goes on actually to apply his suggestions to our specific problem of whether every prescription that is entailed by a prescription in force is also necessarily in force. "Is it not logically possible," he asks,

> . . . that the authority shows great anxiety to make his will effective as far as his manifest prohibitions and orders are concerned, but is completely indifferent . . . as far as the derived prescriptions are concerned?[18]

Von Wright continues,

> [the] answer to the last question is that such an attitude on the part of the norm authority is *not* logically possible. [His emphasis on the "not."][19]

And he explains that this is so roughly because if one manifests anxiety that p be the case, and p entails q, then necessarily one manifests anxiety that q be the case.

As is widely recognized, however, it is a hard fact about the so-called propositional attitudes that they do not preserve the entailment relations holding among their propositional objects. Thus, even though p may entail q, if p or q is complicated enough, someone, S, could well believe, desire, or be anxious about p, but not believe, desire, or be anxious about q. And since one cannot well manifest anxiety which one lacks, it *is* logically possible for S to manifest anxiety concerning p without manifesting anxiety concerning q, regardless of whether p entails q. And so it seems that the "in force" theorist cannot after all derive any support from the suggestions of von Wright.

But that leaves us with our question of what it means for an order to be in force; and it would seem that in answer to this nothing more is forthcoming than the following: "A *derived* order, which *follows from* an order that someone gives, is *in force* (with respect to that someone), either until the order given is satisfied or retracted, or until the derived order is satisfied, or until the juncture of the action it orders is past." But this definition makes use of the notion of "following from" in the logic of orders, a notion simply correlative to that of validity for the logic of orders. This definition would therefore render the concept of "being in force" quite useless for an explication of validity in the logic of orders (and, for similar reasons, quite useless for an explication of validity in the logic of directives generally).

The foregoing discussion leads me to think that though the difficulties presented by the use of satisfaction values may indeed be great, they are not as great as those to which the use of "in force" values gives rise. In fact, it seems clear from what has been said about them that "in force" values are of little avail even in explicating validity for a pure logic of directives (one meant to be applied to arguments containing no statements but only directives). But where "in force" values fail, satisfaction values appear to succeed. In conclusion I shall try to ground this claim by actually providing an explication of validity for pure directive arguments in terms of satisfaction values.

---

[17] *Von Wright* (1963), p. 157.

[18] *Ibid.*

[19] *Ibid.*

Before I can do so, however, our account of directive satisfaction needs to be modified.[20] On our account, a directive is satisfied provided its propositional core is true, where its propositional core is exhausted by someone's making something be true. But that account is not readily applicable to conditional directives. And it seems clear that all directives can be thought of as conditional directives; for those which are ordinarily called "categorical" may be taken to have an implicit, empty condition, such as $p \vee \sim p$.[21] So, the question arises, what do we say of a conditional directive when its condition is false? Is it obeyed or disobeyed? Is it satisfied or violated? My answer, in either case, would be that it is neither, but only "neutral."

Appealing again to our notion of someone's making something be true, we may now say that a conditional directive is "neutral" provided its condition fails to obtain; is "satisfied" provided its condition does obtain and its target or agent does make its instruction true; and is "violated" provided its condition does obtain and its target or agent does not make its instruction true. As an example, "John, close the window if it rains!" is *neutral* provided it does not rain, *satisfied* provided it rains and John makes it true that the window is closed, and *violated* provided it rains and John does not make it true that the window is closed. Given these definitions we may now offer the following as an explication of validity for pure directive arguments.

> (C) A directive argument is valid provided it contains a non-empty subset of conjointly satisfiable premisses such that: (i) if its members are satisfied then necessarily the conclusion is satisfied; and (ii) if the conclusion is violated then necessarily at least one of its members is violated.

This criterion has the desirable feature of not granting validity to such inferences as that from "Don't bite your nails again!" to "If you bite your nails again, bite your fingers off too!" For if the former directive were satisfied, by your not biting your nails again, then necessarily the latter would be neutral, since necessarily its condition would be false; which shows that the argument violates the first condition of (C). Thus we see that (C) does not commit us to the absurd view that the child who disobeys his father in any minor matter, is under orders from his father to do everything.[22]

On the other hand, our criterion also has the apparently undesirable feature of denying validity to the inference from "If it rains, close the window!" to "If it rains and thunders, close the window!"[23] For the former directive may be satisfied, by its raining and your closing the window, while the latter is neutral, because though it rains it does not thunder; which shows that the argument violates the first condition of (C). In view of this, let us now consider some possible alternatives to criterion (C).

In his book on *The Logic of Commands*,[24] Nicholas Rescher proposes a different criterion which when translated into our terminology,[25] reads as follows:

> (C′) A directive argument is valid provided: (i) if the premisses are all satisfied, the conclusion must be non-violated; and (ii) if the premisses are jointly satisfiable, then they are jointly satisfiable with the conclusion.[26]

Concerning this criterion, we may first note that, as Rescher himself points out, it shares the desirable feature which I ascribed to (C), *and lacks* the "apparently undesirable" one. Nevertheless, there is a serious objection to (C′), which we should now consider. But before we do so, and to simplify our discussion, allow me to introduce a few logical symbols. Let us symbolize our conditional directives of the form "John, if p then render it true that q!" as "John, p/q." Since in what follows we may assume John to be our uniform addressee, let us allow "p/q" to be elliptical for "John, p/q." In view of all this, it should be clear that "categorical" directives may now take the form "$p \vee \sim p/q$."

---

[20] The account is given as a definition above.

[21] See *Rescher* (1966), where this idea is first put to use.

[22] The validity of the very plausible inference schema $p, p/q \therefore q$ is presupposed here.

[23] The argument from "Close the window if it rains or snows!" to "Close the window, if it rains!" may be thought to represent an essentially different kind of example from the one we have considered. Nevertheless, it can easily be given the same form. Thus change it to read: "Close the window, if it rains or snows! So, close the window, if both it rains or snows, and it rains!"

[24] See Chapter 7 of *Rescher* (1966).

[25] Given our assumption of a uniform addressee.

[26] See *ibid.*, sect. 7.10, pp. 89–91.

Now, the serious objection to (C′) is that, because of its second clause, it denies validity to the schema:

$$\text{(k)}\quad \begin{array}{l} p/q \\ \underline{p \vee \sim p/\sim r} \\ p \;\&\; r/q; \end{array}$$

while granting validity to the schemata:

$$\text{(l)}\quad \begin{array}{l} p/q \\ \underline{p \vee \sim p/\sim r} \\ p/q \end{array}$$

and

$$\text{(m)}\quad \begin{array}{l} \underline{p/q} \\ p \;\&\; r/q. \end{array}$$

But this comes into conflict with the transitivity of "is validly inferable from."

There is an obvious way of modifying (C′) to meet this objection. Thus we obtain (C″).

> (C″) A directive argument is valid provided there is a non-empty *subset* of the premisses such that: (i) if its members are all satisfied, the conclusion must be non-violated; and (ii) if its members are jointly satisfiable, then they are jointly satisfiable with the conclusion.

Even to this, however, there is a decisive objection.[27] For consider the schema:

$$\text{(n)}\quad \begin{array}{l} \underline{p \vee \sim p/\sim q} \\ q/r \end{array}$$

which turns out to be invalid by criterion (C″). This is a schema the validity of which all of (C)–(C″) are designed to preclude, and about which Rescher writes that it is a "... surely objectionable inference-pattern," which he wishes "... to reject as invalid." But consider the schemata:

$$\text{(o)}\quad \begin{array}{l} \underline{p \vee \sim p/\sim q} \\ p \vee \sim p/\sim q \vee (q \;\&\; r) \end{array}$$

and

$$\text{(p)}\quad \begin{array}{l} \underline{p \vee \sim p/\sim q \vee (q \;\&\; r)} \\ q/r \end{array}$$

These both turn out to be valid on (C″). But in view of the invalidity of (n) we see that (C″) is also in conflict with the transitivity of "is validly inferable from."[28]

Of course, we should not infer from the inadequacy of (C′) and (C″) the adequacy of (C). Indeed (C) is still faced with the "apparently undesirable" feature of denying validity to the inference from "If it rains, close the window!" to "If it rains and thunders, close the window!" Once we conceive of the logic of directives as an organon, however, these appearances may well turn out to be deceptive, since the denial of validity to the inference in question by criterion (C) may then turn out to be a feature of (C) no more than *apparently* undesirable. For the main use of such a logic would then lie in the disclosure, for any given set of directives, of which other directives are such that: (i) they need to be satisfied, in all circumstances, if one is to satisfy each of the directives in the set; and (ii) they need to be non-violated, in all circumstances, if one is not to violate any of the directives in the set. And then it seems quite acceptable that the argument under discussion should be adjudged invalid. Not only is it not the case that its conclusion must be satisfied, in all circumstances, if its premiss is to be satisfied; there is even a circumstance, namely its raining but failing to thunder, where it is possible to satisfy the premiss but impossible to satisfy the conclusion.

University of Western Ontario

Received October 22, 1965

BIBLIOGRAPHY

BERGSTROM, Lars *Imperatives and Ethics* (Stockholm, Stockholm University Press, 1962).

BLACK, Max "Making Something Happen," *Determinism and Freedom*, ed. by Sidney Hook (New York, New York University Press, Washington Square, 1958).

BOHNERT, H. G. "The Semiotic Status of Commands," *Philosophy of Science*, vol. 12 (1945). [Reviewed by F. B. Fitch in *The Journal of Symbolic Logic*, vol. 11 (1946).]

BRAITHWAITE, R. B. Review of Hare's *Language of Morals*, *Mind*, vol. 58 (1954).

---

[27] And, of course, any objection to (C″) is *a fortiori* an objection to (C′).

[28] The disagreement between Rescher and myself concerning the correct criterion of validity should not be allowed to obscure our more fundamental agreement concerning the semantical values to be used in such a criterion.

CASTAÑEDA, Hector N. "Outline of a Theory on the General Logical Structure of the Language of Action," *Theoria*, vol. 26 (1960).

——"A Logico-Metaphysical Inquiry into the Language of Action," in *Morality and the Language of Conduct*, ed. by G. Nakhnikian and H. N. Castañeda (Detroit, Wayne State University Press, 1962).

——"The Logic of Change, Action and Norms," *The Journal of Philosophy*, vol. 62 (1965).

CURRY, Haskell B. *Foundations of Mathematical Logic* (New York, McGraw-Hill Co., 1964).

DUNCAN-JONES, A. E. "Assertions and Commands," *Proceedings of the Aristotelian Society*, vol. 42 (1952).

FISHER, Mark "A System of Deontic-Alethic Modal Logic," *Mind*, vol. 7: (1962). [Reviewed by H. N. Castañeda in *The Journal of Symbolic Logic*, vol. 27 (1962).]

FREGE, Gottlob "Ueber Sinn Und Bedeutung," *Zeitschrift für Philosophie und philosophische Kritik* (N.F.), vol. 100.

——*Grundgesetze der Arithmetik*, vol. 1 (1893).

——"Die Verneinung," *Beiträge zur Philosophie des deutschen Idealismus*, vol. 1 (1919).

HARE, R. M. *The Language of Morals* (Oxford, Oxford University Press, 1961). [See the Braithwaite entry, above.]

HOFSTADTER, A., and McKINSEY, J. C. C. "On the Logic of Imperatives," *Philosophy of Science*, vol. 6 (1939).

LEMMON, E. J. "Deontic Logic and the Logic of Imperatives," *Logique et Analyse*, vol. 8 (1965).

LEONARD, Henry S. "Interrogatives, Imperatives, Truth, Falsity and Lies," *Philosophy of Science*, vol. 26 (1959).

RESCHER, Nicholas *The Logic of Commands* (London, Routledge & Kegan Paul, 1966).

ROSS, Alf "Imperatives and Logic," *Philosophy of Science*, vol. 11 (1944).

SELLARS, Wilfrid "Imperatives, Intentions and the Logic of 'Ought'," *Methodos*, vol. 8 (1956). Reprinted in a revised form in *Morality and the Language of Conduct*, ed. by G. Nakhnikian and H. N. Castañeda (Detroit, Wayne State University Press, 1962).

SOSA, Ernest "Actions and Their Results," *Logique et Analyse*, vol. 8 (1965).

STRAWSON, P. F. *Introduction to Logical Theory* (London, Routledge & Kegan Paul, 1964).

WELLS, Rulon "Frege's Ontology," *The Review of Metaphysics*, vol. 4 (1950–1951).

VON WRIGHT, Georg Henrik *Norms and Action* (London, Routledge & Kegan Paul, 1963). [Reviewed by R. M. Hare in *The Philosophical Quarterly*, vol. 15 (1965).]

American Philosophical Quarterly
Volume 4, Number 1, January 1967

VII. LOGICAL CONNECTION

JON WHEATLEY

I. Introduction

THIS paper concerns itself with the groundwork of the notion of logical connection as it is found in our language. More particularly, it will consider the ways in which two or more statements are connected such that the truth of one is pertinent to the truth of the other as a matter of logic not of contingency. The word "logic" as used here has, of course, nothing to do with an artificial calculus. It is used simply to indicate that the origin of the connections with which I am concerned lies in our language and not in the world.

There are a number of different types of logical connections in our language and, in a short paper, I cannot hope to cover the whole field. I shall, in fact, center attention on one particular type which, I would claim, has never before been successfully isolated. I call it the trident relation. However I shall also, at the end of the paper, have something to say about entailment.

My first task will be to draw a number of rather tiresome distinctions to isolate what is, in my opinion, a very important class of sentences. I call this class the Category of the Logically Opaque.

I shall first, then, draw a distinction and christen it. It is not really a new distinction but I shall draw it in a slightly new sort of way. It is as follows: The word "utterance" will be used to designate an actual utterance produced by some person in some definite context (physical and linguistic), at some definite time, etc.[1] The word "sentence" will be used to designate a set of words in a certain order, complete with punctuation, that constitutes a grammatically well-formed sentence.[2] Thus if two people both say "Get out!" we have two utterances but only one sentence. The phrase "sentence type" will be used to designate the class generated by an incomplete sentence which, when completed, forms a sentence. The most obvious way, but not the only way, to form a sentence type is to take any sentence and substitute variables for one or more words in it, though one can perfectly well form a sentence type by leaving off important punctuation. Examples of sentence types would be "I promise to do x" or "A intends to do x."

The distinction offered here between "sentence" and "sentence type" is not clear in the sense that it is easy to think up intermediate cases where we are not sure what to say. This is entirely acceptable (because, of course, it reflects the facts). Sentences and sentence types do blend into each other, form a spectrum, and no decision procedure is possible with a great many types of cases; it is only important to have two terms available because there are important differences between the ends of the spectrum. Thus, in what follows, I shall often use the term "sentence" to refer to both sentences and sentence types, where differentiation is unnecessary, just to save space. As an historical note, this distinction is more or less that drawn in the past between sentence-tokens and sentence-types, but I shall use the terms "utterance," "sentence," and "sentence type" as being closer to the facts and a good deal more natural.

Just in passing, we should notice that "sentence type" is not as much a philosopher's term as we might at first suppose. When one says "Suppose someone promises something" he is talking in an entirely ordinary way. Yet, in saying this, he is inviting us to consider the sentence type "P promised to do x." The only real point of difference between this usage and talk of sentence types is that by putting named variables into sentence types we have a way to insure that instantiation is uniform over a group of instantiated sentence types.

---

[1] What Austin calls a "speech act" (*vide* J. L. Austin, *How To Do Things With Words* [London, 1962], p. 52). Although Austin invented the term, it is used entirely differently by some writers, e.g., R. Searle, "Meaning and Speech Acts," *The Philosophical Review*, vol. 71 (1962). As it is Austin's term, this would appear to be a piece of presumption.

[2] I have not, of course, defined "grammatically well-formed sentence." This, however, is not a job for philosophers but a job for linguists, and one on which they now seem to be making real progress (especially Chomsky, and linguists of his persuasion). It is also rare, I think, that a genuine philosophical difficulty turns on whether or not a given set of words is a grammatically well-formed sentence or not. That is, it is rare that we are genuinely in doubt as to whether a putative sentence is grammatically well-formed or not.

II. Contradiction and Its Relatives

Strawson, in his *Introduction to Logical Theory*, wrote as follows:

> Why bother to avoid inconsistency? What is wrong with contradicting yourself? There is nothing morally wrong with it. It may not even be entirely pointless. Suppose a man sets out to walk to a certain place; but, when he gets halfway there, turns round and comes back again. He may, after all, have wanted only exercise. But from the point of view of a change of position, it is as if he had never set out. And so a man who contradicts himself may have succeeded in exercising his vocal chords. But from the point of view of imparting information, of communicating facts (or falsehoods) it is as if he had never opened his mouth. He utters words, but he does not say anything. Or he might be compared with a man who makes as if to give something away and then takes it back again. He arouses expectations which he does not fulfill; and this may have been his purpose. Similarly, it may have been the purpose of a man who contradicts himself just to create puzzlement. The point is that the standard purpose of speech, the intention to communicate something, is frustrated by self-contradiction. Contradicting oneself is like writing something down and then erasing it, or putting a line through it. A contradiction cancels itself and leaves nothing.[3]

The idea of sentences or types of sentences which have no use or, as Strawson has it, cannot fulfill the standard purposes of speech, do not communicate something, is a very interesting one. But it does not do quite the job Strawson wants done here because far more types of sentences than contradictions fail in the required way. For instance, shortly after Strawson's book came out, Hamlyn was arguing in *Mind* that there are no analytic statements. His reasoning, in outline, was very simple: putative analytic statements, he said, have no use, cannot be used to communicate anything. (So—his argument went on—anything that is analytic is not a statement and there are, as a consequence, no analytic statements.) If Hamlyn is right in saying that analytic sentences have no use, then the principal feature which Strawson ascribes to contradictions, that they can communicate nothing, is shared by all analytic statements; and clearly that won't do. Nor do we have to depend on Hamlyn here. To say (instantiated) "I assert that *s* but I do not believe *s*" or "I promise to do *x*, but I don't intend to do *x*" equally have no use, or only very special types of uses (of the types that contradictions have). None of this fits Strawson's thesis.

I do not think there is much doubt that Strawson's account of contradiction, as I have quoted it, is incomplete. But that does not mean it is dead wrong. In the notion of a sentence having no use or function, or only a very odd use or function, or not fulfilling the standard purposes of speech, lies, I believe, the solution to many of our problems about logical connection. What Strawson has done, I shall claim, is given us an imperfectly explicit definition of a far larger class of sentences than the class of contradictions, though his class contains contradictions as a subclass. I call the whole class the Category of the Logically Opaque.

III. Logical Opacity

I would like to define a logically opaque sentence as one which is not nonsense and yet has no use whatever, that is, a sentence which could never attain to the level of an utterance. This, however, would not do at all. I doubt if there is a meaningful sentence of the English language such that a person of sufficient ingenuity could not think up some context in which it would be plausible to suppose it uttered.[4] Nor is there a way out, the way Strawson thinks there is, by talking about the "standard purposes of speech" or "communicating something" for these phrases, though they indicate the root idea I am interested in, are just too vague to carry the weight they must carry.

The way I shall choose, which is I am afraid a long and tiresome way, is to say that a logically opaque sentence is one which, if it has any use at all, has only certain peculiar sorts of uses. This notion requires a good deal of fairly careful spade-work.

First we must set up conditions of utterance for sentences such that they will count as single utterances. The required conditions are as follows: that an utterance, to be an utterance, must be given consecutively and the audience to it must remain constant. The necessity for this condition must be seen as follows: suppose I am giving evidence at some tribunal with a confidant standing beside me. I say, loudly to the whole group, "I promise to tell the whole truth and nothing but the truth" and then I add, *sotto voce*, to my confidant, "but I'm not going to." Or I say to a group

---

[3] P. F. Strawson, *Introduction to Logical Theory* (London, 1952), pp. 2–3.

[4] *Vide* A. Stroll, "Uses of Analytic Statements in Ordinary Discourse," *Mind*, vol. 66 (1957), pp. 541–542 (a comment on Hamlyn's paper on analytic statements).

"I shall go to the meeting this evening" and add, *sotto voce*, to a confidant (or to myself, inwardly) "but I won't." In both these cases the whole utterance "I promise to tell the whole truth and nothing but the truth—but I'm not going to," "I shall go to the meeting this evening—but I won't," suffer from the sort of faults Strawson had in mind when he spoke of not fulfilling the standard purposes of speech. Yet, because of the peculiar conditions of their utterance, they function perfectly well on the sorts of occasions cited. It is quite clear that something untoward is going on here and that to think of both parts of the utterance as being parts of some *one* utterance is misleading. However, there is no need to delay over precisely how to describe such cases so long as a clear condition for something being one utterance is established; this I have done above.

Second, we must give a list of logically opaque uses of sentences of certain types; this list constitutes the definition of logically opaque uses. However, to state a contention I shall later defend, the definition is not arbitrary.

Logically opaque uses are as follows:

1. The use of sentences (the utterance of sentences) to exercise one's vocal chords, make a noise (say, to let someone know one is present), see how some word or group of words sound (the way some women try out the sound of what they hope will become their married names), and so on.

2. The use of sentences as part of games, rituals, ceremonies, incantations, chants, songs, prayers, advertising slogans, and so on.

3. The use of sentences to contradict oneself (all in one utterance) or avoid saying anything; the use of sentences to confuse, distract, perplex, muddle, or puzzle people.

4. The use of sentences as examples (i.e., when they are mentioned rather than used in a tighter sense of "use"). This type of use occurs most often in grammar books, spelling lessons, works in philosophy, philology, linguistics, phonology, etc.

5. The use of sentences which contain no interior quotation marks to explain or exemplify or give the meanings of one or more of the words or constructions contained in them.

It will be noticed that these types of use are not exclusive, that is, a given use of a sentence often falls under one or more of these types. Equally, there will be awkward cases where we have difficulty deciding whether the use in question falls into one or more of these categories or not. This last is inevitable; language is not so inflexible that it is ever possible to have important categories without borderline cases.

The property that these uses hold in common is this: that in all of them more is required to understand what is going on than just a thorough knowledge of the English language, i.e., the semantic content of what is said is insufficient to fully understand what is going on, and a knowledge of the language is insufficient to indicate this. Thus when one sees the statement "I am in pain," a mere knowledge of the language is insufficient to know what is going on (who uttered it is crucial), but the language indicates this. That is, anyone who knew the language would realize this. But when a child of six sings a song containing statements about her lover or her children we have to know more than just English to realize what is going on. Similarly, when a philosopher uses the statement "All crows are white" as an example of something or another we do not understand what is going on by looking at the statement alone. Again, when someone says "All bachelors are unmarried" we have to realize that the statement is not about bachelors but, in so far as it is about anything, it is about the meaning of the word "bachelor." Or, if this last remark is objectionable, it is at least not about bachelors the way the statement "All students are rebels" is about students.

I must emphasize that the description given above is far too vague and approximate to constitute a definition, even with the illustrations, and it is not offered as a definition (the list gives the definition); it is offered as an explicative description only.

Finally, for purposes of exclusion, one must have available the notion of nonsense. It is not well defined and must remain so. I can only say about it that, as used here, the term "nonsense" is taken to cover only those sentences which, in the most literal sense, would ordinarily be called nonsense. That is, I wish to say of a sentence that it is nonsense when it is grammatically well formed (being a defining condition of a sentence) and when the words in that order appear to make no sense at all.

A nonsensical sentence is extremely hard to invent, and when one actually occurs it frequently involves neologisms or phony morphological variants. Here, however, is an example which was not intended to be nonsense: "The accent which was on the concern for Being, opens, exteriorizes, expresses itself" (from *What is Phenomenology* by Pierre Thevenaz). And here is an example, without neologisms, adapted from Chomsky: "Humerous green ideas sleep furiously." I am more doubtful about the Chomsky example than about the other; that is, I cannot help wondering whether someone more ingenious than I could not think up some situation in which this sentence might not be used to make perfectly good sense. The classic example of "Saturday is in bed" is clearly of this type: as we all know, Friday night is drunk, Saturday's in bed, Sunday's in church, and Monday is back at the old grind again. That is, I doubt whether there is any grammatically well-formed sentence of the language which is nonsense in the required sense. However, this suspicion, right or wrong, makes no difference to the use I make of the notion below.

IV. THE CATEGORY OF THE LOGICALLY OPAQUE

The category of the Logically Opaque is defined over sentences. The definition is as follows: those sentences which are (a) not nonsense and (b) either have no use or can *only* have logically opaque uses, constitute the Category of the Logically Opaque. It is not the case that sentences can sometimes belong to the Category of the Logically Opaque and sometimes not, though some sentences undoubtedly do have some logically opaque uses and other uses as well; it is that only those sentences which have *no* use or *no* use other than logically opaque uses, and are not nonsense, constitute the Category of the Logically Opaque.

I wish to divide the Category of the Logically Opaque as follows: The sentence "Circles are square" falls into the Category of the Logically Opaque. So, therefore, does the sentence "Circles are square and most men wear trousers." However, what makes the sentence, "Circles are square and most men wear trousers" Logically Opaque is the Logical Opacity of the sentence "Circles are square." When this situation arises, i.e., when a sentence is Logically Opaque by virtue only of containing a sentence which is Logically Opaque, I wish to say that it falls into the subcategory of "consequent Logical Opacity." When I wish specifically to refer to Logical Opacity of both subtypes I shall use the phrase "all type Logical Opacity." I shall use "Logically Opaque," without prefix, for pure Logical Opacity.

The reason why the Category of the Logically Opaque is worth all this trouble to define is that it contains a great number of philosophically puzzling sentences. For instance, sentences which are analytic and ones which are contradictory fall into the Category of all-type Logical Opacity; and they are of most interest when they fall into the Category of pure Logical Opacity. Sentences involving entailment relations are more complicated: when we have a sentence of the type "*p* entails *q*," it falls into the Category of the Logically Opaque when its utterance would constitute a true statement. There are a good many further types of sentences within the Category of the Logically Opaque of which I shall give only a sampling. There are those types of sentences, first noticed by Moore, such as "*p* but I don't believe it," "I promise to do *x* but I don't intend to," "I know that *p*, but I don't believe it" and so on. There are Strawson's cases of presupposition: "There is no present King of France and the King of France is wise," "John has no children and they are all bald." There is also the highly interesting category of self-defeating sentences, i.e., sentences which, if uttered, would, by the fact of their utterance, defeat their content. In the simplest instance these are personal avowals which the fact of avowal invalidates, though there are more types than this. Examples are: "I am asleep," "I am dreaming," "I am lying (when I say this)," "I do not exist." Philosophers we could all name have made large reputations just by asserting a statement where the sentence formed by negating the statement falls into the Category of the Logically Opaque. The distinction between those sentence types where every instantiation falls into the Category of the Logically Opaque and those where only some instantiations fall into that Category will be found to be crucial. These are not, however, the only types of sentences in the Category of the Logically Opaque but they represent a fair sampling.

It should be noticed that the Category of the Logically Opaque, as I have defined it, includes only sentences. However, it is perfectly possible, and highly desirable in a work of greater compass than this, to define a wider category of the Logically Opaque to include utterances and phrases. The Category of Logically Opaque utterances (which are not just the uttering of Logically Opaque sentences) would include such utterances as "You go that way" said while pointing two ways at once, or "Pick up that pipe" said while pointing at a rose, and so on. The Category of Logically Opaque phrases would include, in the most obvious instance, adjective-noun constellations which, used together, cannot be used to fulfill the standard purposes of speech, e.g., "deliberate yawn," "unintentional murder," and so on. I mention these additional possibilities only to give an idea of how the theory of logical opacity could be extended and of how doing so includes a great many of the most interesting areas where logic and language come into interesting interaction.

V. Preliminary Description of the Trident Relation

I shall now introduce the trident relation informally by considering some examples.

Consider the sentence type:

A promised to do *x*

This does not entail

A will do *x*

nor yet

A will try to do *x*

nor yet

A intends to do *x*.

Yet there are logical connections running from "*A* promised to do *x*" to all these sentences. These logical connections do not derive from the meaning of the word "promise," ordinarily conceived, but from the use of sentences of the type "I promise to do *x*" when actually uttered. Similarly, the sentence

A said, with a straight face, not on the stage, etc., '*s*'

(where *s* is a statement variable) does not entail

A believes *s*.

It is nevertheless a fact that part of the primary purpose of saying something is to give an indication that one believes it. Again, the sentence,

A groaned and cried

does not entail

A is in pain

but there is certainly a connection.

It is, of course, not news to note the logical relationships between these various types of sentences. Wittgenstein pointed out that there is a non-contingent relationship between pains and their expression; Austin said that there is a different type of implication relationship between *A*-saying-*s* and *A*-believing-*s* but his one attempt to characterize it failed;[5] Hare makes an untenable claim (or implies that such a claim is correct) that there is an entailment relation between how we act and what moral principles we hold (accurately, of course, an entailment relation between statements as to what moral principles we hold and statements of how we will act). Though this claim is clearly false there is, equally clearly, a logical relationship between them. What is new in what follows is that I attempt to give a definition and an analysis of this relationship.

VI. Definition of the Trident Relation

The word "trident" is deliberately neutral; it is derived, in fact, from the symbol for it, which is '∃'. The verb from "trident" is "to tride" so that "$p \exists q$" reads "*p* trides *q*." Although the trident relation is a logical relationship it is not truth functional and cannot be defined into the propositional or predicate calculi; but this is hardly surprising since the same remarks may be made about entailment. The trident relation holds between sentence types, not sentences. It is defined below.

The following notation is used: the letters *p* and *q* are used for sentence types. An instantiation of a sentence type is that sentence generated by substituting allowable substituends for the variables in the sentence type; an instantiation of *p* is written p' and the same instantiation (i.e., same values for the variables) for *q* is written q'. Whenever I speak of an instantiation of *p* and *q* (or *p* and not-*q*) I always mean the same instantiation in both cases. Then, consider this situation: that whenever an instantiation of *p* is true, then the same instantiation of *q* would be false (call it the defeating condition). We can now define the trident relation as follows:

$p \exists q$ if and only if

(a) the occurrence of the defeating condition would change, or indicate there had been a change in, our language in the relevant area.

(b) There is at least one instantiation for *p* and *q* such that the sentence "p' but (and yet) not-q'" falls into the Category of the Logically Opaque.

The notion of language change, which occurs in this definition, appears to be a difficult one. In particular, it appears to be systematically riddled with difficulties over borderline cases. Suppose we know, as in many instances we do know, that the language has changed. Such a change, unless it is a matter of definition, will take place gradually and the actual change will be undatable. Thus, through the period of change (which, in actual historical situations, is often a century or more), the question "Has language changed?" will be, to all intents and purposes, unanswerable; and if it is answered the answer will doubtless be in part a prediction of how the language will be in the future. However, to note this is *not* to note a difficulty with the definition of the trident relation. To ask, "Granting language is changing, when can one say that language has changed?" is to ask a very different question from "Postulating a *dramatic* change in speech patterns, would that change, or indicate there had been a change in, the language?"; it is with the second sort of question that the trident relation is concerned. To take an actual example, the word "presently" has changed in meaning since Shakespeare's day: then it meant what "immediately" now means; now it means what "soon" now means. To date this change, or say what normal speech patterns would have to be different before we could say correctly "The word 'presently' has now changed meaning," is doubtless impossible or, if done, would be in large part arbitrary. But applying the definition of the trident relation is in no sense concerned with trying to answer this sort of question. It is for this reason that giving general criteria of when we would say language has changed is not necessary to explicate the trident relation; we can be content with the sort of very general and diffuse criteria we already have which allow us to say *with complete certainty* that "presently" has changed meaning since Shakespeare's day.

---

[5] See my "Austin on Implication and Entailment," *Philosophical Studies*, vol. 15 (1964).

VII. Discussion of the Trident Relationship

I shall first give two examples of the trident relation and then go on to discuss some of its more technical properties, in particular say what has to be said about quantification and the fact that the relation only holds between sentence types.

Consider these two examples of the trident relation:

A asserts $s \exists A$ believes s (1)

A says, "I promise to do x" $\exists A$ will do x (2)

In the case of (1), the relation states, what we all know, that the primary function of asserting is to state, at the least, one's beliefs. Equally, as we all know, people lie, so asserting is not tied *via* an entailment relation to believing. However, if in general people started asserting things where their beliefs were in no sense involved (i.e., if the defeating condition obtained), then language would have changed. Not necessarily changed for the worse, or changed so as to make the language useless, but there would have been a change. It is this which makes the relationship between asserting and believing logical and not factual and which makes the statement of the trident relationship true.

In the case of (2) above, "A says, 'I promise to do x' $\exists A$ will do x," the statement of the trident relationship is true because it is the function of promises to bind us to do certain things in the future, though this does not deny that promises can be broken without illogicality. Equally, if people said "I promise to do x" (instantiated) from time to time but that their saying this was completely independent of their conduct, our language would have changed.

A further note is perhaps in order here. The definition of the trident relation is not easy and it may well be that the trident relation can be better defined. However, it is fruitless to hope for a definition which provides an easy decision procedure in all cases. The statement that a trident relationship holds between p and q is, in most cases, a substantial philosophical position and, like a similar entailment statement, must be debated in detail like any other substantial philosophical position. By offering a definition of the trident relation which catches what is important here, as I hope I have done, I have offered a statement of what the assertion of a trident relationship involves; that a good deal of debate is still possible as to whether any given statement involving the trident relationship is correct is a monument to the complexity of our language, and relations within it, not of itself a fault in the definition.

VIII. Technical Properties of the Trident Relationship

I shall now discuss some more technical aspects of the trident relationship as I have defined it.

The trident relation holds between sentence types or, in jargon, unquantified open sentences. It does not necessarily hold between subclasses of sentence types between which it does hold. That is, the statement

Jones asserts $s \exists$ Jones believes s

is not correct, for the constant conjunction of Jones asserting something while not believing it (i.e., the defeating condition obtaining) shows nothing about the language but only that Jones is a chronic liar. Similarly, if the range of some variable in the sentence type is contracted the result may be a false statement. Thus if, in the trident statement,

A asserts $s \exists A$ believes s,

the range of s is limited to statements about Santa Clause or sex morals, the trident relationship no longer holds. That is, the constant conjunction concerned would show nothing about our language but only something about the lies we habitually tell in our society. It is for these reasons that statements of the trident relation cannot be quantified. The point is that the universal quantifier, which is the appropriate one to consider, is bound up with its instantiation rules, which do not hold for trident statements. The alternative appears to be to invent a special quantifier to read, roughly, "generally." However, such a quantifier makes it sound as if the trident relation expresses a statistical correlation, which is not the case, so there seems no advantage to the invention. I prefer, instead, to leave the statements unquantified. One then reads trident statements as ordinary statements in English (except, of course, that they contain a well defined technical term) connecting types of sentences.

It might be felt that even if what I have said above is granted, it is still the case that if an entailment relation holds between sentence types then this itself entails that a trident relationship also holds between the two sentence types. This is true as a statement couched in ordinary language. Thus

(A is a father) entails (A is a male)

and it follows from this that

A is a father $\exists$ A is a male.

However, once the entailment statement is put into symbolic terms with the universal quantifier and material implication, difficulties can arise. But as I am in no sense trying to invent a new logical calculus, this problem need not concern us. The connection between entailment and the trident relation is discussed in Sect. IX.

The trident relation also has the following properties: It is reflexive, it has none of the properties of symmetry (i.e., is not symmetric, asymmetric, or antisymmetric, except the last by very special definition) and it is transitive. From the point of view of the philosophy of language, what is of interest in the trident relation is that it states, or signals, a fact about language; however, it has a real practical application as a quasi-rule-of-inference. That is, heavily modified forms of the *Modus Ponens* and *Modus Tollens* hold. For *Modus Ponens*, the modified form is as follows: if p' and q' were appropriate instantiations of the sentence types p and q, and if $p \exists q$ is true, and if p' is true then, all else being equal, q' may be inferred:

i.e. $$\frac{p' \qquad p \exists q}{*q'}$$

where '*' is read "all else being equal" or something similar ("probably" is to be avoided here). Under the same conditions, there is a *Modus Tollens* as follows:

$$\frac{\bar{q} \qquad p \exists q}{*\bar{p}'}$$

I must emphasize that these practical applications are in no sense something invented by me. We see them at work in this sort of way. Someone says

Jones is certain to be here

and, challenged ("Why?"), gives a reason to support his statement, as follows:

He promised to come.

Without information to the contrary, e.g., that Jones habitually promises without performing or that he has, in the interim, been run over by a car, the statement "He promised to come" is a reason (though not a conclusive reason) for saying "He is certain to be here."

However, these practical applications are not my concern now. The point is that there are all sorts of ways to reach a valid conclusion of the form "All else being equal, p," namely all those occasions when we give non-conclusive reasons for p, where there is no logical relationship involved. The importance of the trident relation lies in the fact that it is a logical relationship.

IX. Definition of Entailment

If I have been successful in my definition of the trident relation, a definition of entailment drops out with no further effort. It is given below.

The sign '$\Rightarrow$' will be used for "entails." As before, p and q are sentence types. The defeating condition (the same as for the trident definition) is as follows: that whenever an instantiation of p is true, the same instantiation of q would be false. We can now define entailment as follows:

$p \Rightarrow q$ if and only if

(a) the occurrence of the defeating condition would change, or indicate there had been a change in, our language in the relevant area;

(b) Every instantiation for p and q is such that the sentence "p' but (and yet) not-q'" falls into the Category of the Logically Opaque.

The above definition of entailment is not the simplest possible in the terms I have been using, but I give it in this way to bring out the relationship between the trident and the entailment relationships. There is no need to define entailment over sentences as well as sentence types for obvious reasons.

If I have done my work well, this definition of entailment exactly coincides with our intuitive notions of entailment and there are no paradoxes.

University of California, Santa Barbara

Received November 15, 1965

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 1, January 1967

VIII. IS CHOICE DETERMINED BY THE "STRONGEST MOTIVE"?

REM B. EDWARDS

MOST determinists from Spinoza and Jonathan Edwards to those of today have argued that human choice is always determined by the strongest motive;[1] and if their libertarian opponents are to defend their own position, they must show, among other things, that there is something wrong with this contention. Many introductory textbooks in psychology and many philosophers accept uncritically the determinist position, but has not the question been closed prematurely? In this paper I shall deal with the question as it has been discussed by Sir David Ross and C. A. Campbell.[2]

It should be noted at the outset that the major dispute between the determinist and the libertarian is over whether it is the strongest motive which determines *choice*, and not over whether it is the strongest motive which determines *action*. "Choice" here is always used to refer to psychological selection, whereas "action" always refers to physical activity. "Motive" is understood very broadly as any desire, interest, inclination. The disputants could perhaps agree that the "strongest motive" determines action, but they do not agree that the "strongest motive" determines choice. The determinist argues that choices are always determined by the strongest desire which is itself the ultimate outcome of antecedents in the character and environment of the individual. In contrast to this, the libertarian *should* argue that the actual situation is just the reverse, that in those situations where choices occur it is the choice which determines what the "strongest motive" is to be. Perhaps the strongest motive then determines action, but choice determines the strongest motive.

The question which most needs to be settled in the present dispute is: Which comes first, the choice or the strongest motive? What is the order of succession involved? Does the choice determine which motive is to be strongest, or does the strongest motive determine which choice is to be made? For the determinist, the succession of events always seems to be something like this: first, there is a set of competing motives of relatively equal strength, then the predominance of one of these which becomes "strongest," then a choice determined by the strongest motive, and finally where the choice issues in physical consequences there is an action. In contrast, the libertarian offers this order of succession: first there is a set of competing motives, then an act of choice which causes one of the competing motives to predominate, then an action which is determined by the predominating motive. An attempt to resolve the question of the order of succession involved will be made later.

Before attempting to resolve this dispute, considerable attention needs to be given to the notion of "strongest motive." This notion was originally based upon a model taken from physics and applied to psychology, but in our day various attempts have been made to modify the older position. The older attempt to construe human motivation by analogy with mechanistic physics is expressed in Schopenhauer's statement that conflict among motives "puts the will in the same situation as that of the body on which different forces act in opposite directions, until finally the decidedly strongest motive drives the others from the field and determines the will."[3] Twentieth century determinists such as Sir David Ross have attempted to divorce the "strongest motive" pattern of explanation from the physical models upon which it was originally based, and they have suggested their own theories do not share in the crudeness and inadequacies of the older formulations. Let us look at some of the points at which the older mechanistic analogy has been found to break down, even by the determinists themselves.

---

[1] For example see: R. H. M. Elwes (ed.), *The Chief Works of Spinoza*, vol. II (New York, Dover Publications, 1955), pp. 134, 194–195; Jonathan Edwards, *Freedom of the Will* (New Haven, Yale University Press, 1957), p. 141; Arthur Schopenhauer, *Essay on the Freedom of the Will* (New York, The Liberal Arts Press, 1960), p. 37; Hastings Rashdall, *The Theory of Good and Evil*, vol. II (London, Oxford University Press, 1938), p. 306; W. D. Ross, *Foundations of Ethics* (Oxford, The Clarendon Press, 1939), pp. 229, 230; Norman L. Munn, *Psychology* (Boston, Houghton Mifflin Co., 1946), p. 225.

[2] W. D. Ross, *Foundations of Ethics* (*op. cit.*), and C. A. Campbell, *On Selfhood and Godhood* (London, George Allen & Unwin Ltd., 1957).

[3] Schopenhauer, *Essay on the Freedom of the Will* (*op. cit.*), p. 37.

It must be acknowledged at the outset that there is a certain amount of *prima facie* appropriateness about the analogy between "strongest motive" in psychology and "strongest force" in classical physics. Certainly it is true to experience that desires or motives do differ in "intensity" or "strength." Particular desires differ in intensity at *various times*. The desire for sexual expression is certainly more intense at some times than at others, and similarly the desires for food, drink, exercise, companionship, etc. Furthermore, it does make sense to *compare* different desires with respect to their strength. Under certain embarrassing conditions, the desire to relieve the pressure on the urinary bladder may become an all-consuming passion, which drowns out more reputable desires to pay attention to the lecture or listen to the symphony! Examples of such conflict are easy to find.

Although there is some obvious appropriateness about the analogy in question between "strongest motive" and "strongest force," some attention must now be given to the respects in which the analogy fails to hold. Ross has himself suggested two such points, but there are three which will be here discussed. First, there are important differences in the ways in which "strongest motives" and "strongest forces" are detected and measured. Next, there is no psychological theorem comparable to the "parallelogram of forces" as this appears in physics. Finally, it can be argued that if the likeness between motive-causes and force-causes were complete, there would be no need for an "act of choice." Choosing would be made superfluous.

I. Definition and Measurement of "Strongest Motive"

The analogy between "strongest motive" and "strongest force" breaks down at the point of measurement. A "behavioristic" test to decide which force is "strongest" is universally acceptable, but a behavioristic test for the motive which is strongest seems only to beg the question whether choice is always determined by the strongest motive. Also, there are fundamental ambiguities in the whole notion of "strongest motive" which have their origin in alternative criteria which are provided for detecting the strongest motive. Does "strongest motive" mean "motive which is felt most intensely," in which case it would have to be measured by introspection; or does it mean "motive which in fact determines action," in which case it could be measured or detected through behavioristic criteria?

Determinists have favored behavioristic tests for deciding which motive is to qualify as "strongest." Ross for instance states that "it is only by our success or failure to do the act that we can discover whether the desire was strong enough."[4] Other determinists have argued that the only way we can tell which motive is strongest is to wait and see which one does, in fact issue in action. For the determinists, "strongest motive" thus means "motive which was successful in bringing about an observable action." An experimental psychologist might set up an experiment to determine the "strength" of motives in rats using the test suggested by the determinist. A maze might be constructed which would require the rats to make "choices" between alternative paths. The "strongest motive" would be the motive which actually determines the rat to proceed along one path rather than another. Applied to the human species, the test is basically the same. The "strongest motive" is thus merely the "motive which in fact prevails."

What is wrong with this experimental criterion for deciding which motive is "strongest"? For one thing, as thus far stated, it is always an *ex post facto* criterion, whereas it does seem possible for us to decide what we desire most in some instances far in advance of acting upon those desires. When I dine in a restaurant, I am usually able to decide which meal I most desire to eat *before* I give the waitress my order—I do not have to wait to see what I order before I know which I most want! This sort of explanation has the cart before the horse! We can know in advance that we will order steak rather than franks, that our desire for steak will prevail *because* we already know it to be strongest in some non-experimental sense, and the experimental criterion is not needed to tell us that it is the strongest. It will prevail because it is strongest; its being "strongest" does not mean that "it will prevail."

There are still other difficulties with experimental tests for deciding which motive of a set of competing motives is strongest. Some of my strongest desires do not issue in immediate action to be thus tested, and some never in observable action at all. Some are divorced in time from the moment in which they actually issue in publicly observable behavior, as when I resolve in August to give up cigar-smoking next Lent. Some never issue in observable action at all, as when I desire to daydream about my wife rather than about taking a rocket ride into outer space.

---

[4] Ross, *op. cit.*, p. 50; See also p. 222.

The most serious difficulty in treating "strongest motive" as "motive which in fact prevails" is that this definition begs the whole question as far as the controversy between determinism and libertarianism is concerned. Does the strongest motive determine the *choice* that issues in the action or the unobservable consequences that follow in my imagination? Or are there at least some cases in which choice determines what the strongest motive is to be? Does the strongest motive appear before or after the act of choice has been made? The behavioristic criterion will not answer this most important question for us. The observable fact that a given motive does in fact prevail does not decide at all between the determinist's contention that it is the desire as *already* strongest which determines the choice of an end, and the libertarian contention that the desire itself might have become "strongest" because it was *supplemented* by an act of choice to prevail over its competitors. Both sides can agree that the motive which in fact prevails is the motive which in fact prevails, but the determinists cannot define "strongest motive" this way without begging the very question on which he supposes himself to be in disagreement with the libertarian. In the discussion to follow of Ross's theory that choice is determined by the "universe of desire," attention will be given to the second criterion which in fact usually operates in deciding which motive is strongest, according to which "strongest motive" is "the motive which is introspectively felt most intensely."

II. "STRONGEST MOTIVE" AND THE "PARALLELOGRAM OF FORCES"

A second point of difference in the analogy between "strongest force" in physics and "strongest motive" in psychology is one which Ross himself has admirably developed. There is no psychological theorem comparable to the "parallelogram of forces" in physics. A relatively brief discussion of this point is all that is required. Physical forces have a resultant which is the compound result of both acting together, but in psychology, choices which are made on the basis of the "strongest motive" will not be the compound results of the strongest motive *plus* the weaker motives which were competing with it for dominance (so Ross points out). The choice is simply the result of the strongest motive, and the weaker motives have no results whatsoever. Ross tells us that people simply make up their minds what they want most, recognize that some of the things which they want are incompatible with what they want most, and suppress as far as possible their subordinate desires. Even if the subordinate desires still persist, they do so as "mere longings" which do not modify their actions. Their action is not the compound result of the strongest plus weaker desires, but of the strongest desire alone. Weaker motives in psychology are ineffectual, but weaker forces in physics do partly determine the physical resultant. Ross points out that if the parallelogram of forces held in psychology as it does in physics, "our action would be an attempt to get to some extent each and all of the many things we happen to be desiring."[5] Obviously this is not the case. We do not attempt to go partly to the theater and partly to the lecture, and end up somewhere in between! We choose the theater, and the desire to go to the lecture does not result in any action at all. Ross brings out the contrast between "strongest motive" and "strongest force" most effectively in asserting:

> This, it seems to me, is the great difference between physical and mental causation, that in the latter there is no law of the composition of all the forces concerned, but some of the forces concerned are, by an act of choice, deprived of any effect on action.[6]

With his discussion of the breakdown of the physical-psychological analogy, we can agree; but there is still some question as to whether or not he is correct in thinking that choices are themselves determined by the strongest motive.

III. "STRONGEST MOTIVE" AND CHOICE

A final important point of difference between the role played by strongest motive in psychology and strongest force in physics is that there is nothing corresponding to "choice" in the latter. Ross himself also notices this, but it is doubtful that he has himself taken the significance of this fully into account. If the analogy between strongest motive and strongest force were very strict, there would be no need for anything like choice since the strongest motive would result *directly* in action, without having to pass first through an act of choice in order to become effective. He who tries to force the likeness between strongest motive and strongest force is required either to deny completely that there is such a thing as an act of choice or else to make choice superfluous in bringing about observable results. It must be said in Ross's favor that he neither denies that identifiable choices occur nor does he mistake them for some other kind of thing which they are not, such as motives.[7] And though he professes not to make choosing superfluous, unfortunately on this he seems to make the same mistake himself which he criticizes in his deterministic predecessors, as will be shown in discussing his theory that choices are determined by the universe of our desire.

---

[5] *Ibid.*, p. 229.

[6] *Ibid.*, pp. 229–230.

IV. CHOICE AND THE "UNIVERSE OF DESIRE"

The problem of finding an adequate definition and test for "strongest motive" is still unresolved at this point. The high initial plausibility of the formula that choice is determined by the strongest motive we have seen to be derived from an assumed closeness of likeness between strongest force in physics and strongest motive in psychology, but this analogy has broken down already at several points. The analogy is very important to the determinist, and it is imperative that the libertarian expose its weakness. If there is to be freedom in the libertarian sense, choice must not be determined by the strongest *set* of motives, no matter whether the members of this set be conscious, unconscious, habitual, dispositional, sporadic, or however they may come.

Sir David Ross maintains that his own theory that choices are often determined by the strongest set of my desires, rather than by the strongest individual desire, is regarded by him as a refinement upon older deterministic interpretations of the relation between motivation and choice. Further, he regards it as having the peculiar virtue of not making choosing superfluous as older theories had done. Let us look at his claims in more detail.

Previously it was suggested that it might be possible to define the notion of "strongest motive" as "motive which is felt most intensely," a definition which would not suffer from the weaknesses of the behavioristic definition as "motive which in fact prevails." In saying that my desire to attend the jazz concert rather than the poetry reading is my strongest motive, I seem to mean that I directly feel a much more intense attraction toward the former than toward the latter. The introspective test is in many cases an adequate test for deciding which motive is strongest. Now, it should be noted that Ross's theory that my choice is sometimes determined by my "universe of desire" rather than my strongest single motive is a rejection of the older and simpler theory that choice is usually determined by the single desire which I feel most intensely. (Ross himself applies the introspective test, though he also comes out for the behavioristic test, as indicated above.)

Ross holds that there are *some* cases in which I act against my strongest single desire, the one which I feel most intensely. In such cases, my choice is still determined, however. It is determined by the "universe of my desire." He writes that "it is the universe of my desires that determines my action, and not the strongest single desire."[8] A single interest which is originally felt to be strongest is seen to "harmonize less well" with the universe of my interests, and in such cases it is the universe of my interest which prevails. This does seem to be in accord with the facts, and examples could easily be cited. It must be noted that Ross uses several phrases to refer to the fact that choice is sometimes determined by the "universe of my desires." In addition to this expression, he sometimes refers to "the strongest mass of desire"[9] and sometimes to the strongest "set of motives."[10] This latter term seems to say more precisely what Ross has in mind, and it will be used henceforth.

Ross believes that one of the chief virtues of his new theory is that it allows choosing to have a real function of some sort and does not make it superfluous as did the older determinisms which were too close to the analogy with physics. What is meant by making choosing superfluous? What does choice do that is not superfluous? Why are choices ever needed anyway? What do they accomplish that is not achieved by motives? Is the process of deliberating and deciding related to motivation, in Broad's words, as "a mere idle accompaniment which can only give a formal recognition to a *fait accompli*, as the King does when he gives his assent to an Act of Parliament"?[11] Does the determinist's theory that choices are always determined by the strongest motive *or set of* motives make choosing superfluous, or does it not? Can the determinist explain why we ever have to make choices anyway?

---

[7] *Ibid.*, p. 126.

[8] *Ibid.*, p. 229.

[9] *Ibid.*, p. 230.

[10] *Ibid.*, p. 234.

Ross himself raises the question whether choice is made superfluous by determinism. He asks:

> If action is described as following upon the strongest desire, is not resolution or choosing or deciding made completely nugatory? Would not the same act follow if the step called choosing were entirely omitted?[12]

The answer which Ross gives to his own question is not at all clear, unfortunately. He claims that the answer to his question is contained in what he had already written about the theory that choice is determined by the strongest "set of motives" rather than the strongest single motive. This is precisely the difficulty that his new theory is designed to meet, but it is difficult to see any relevance. Ross explains that choice is determined "not by the strength of the isolated desires as they were before the process of deliberation, but by the strength of the appeal which one act, with all that it is expected to involve, makes on us, as compared with the appeal which the alternatives make."[13] He concluded that "what is resolved on may be, and often is, very different from what would have been done if deliberation and choice had not intervened."[14] Now, Ross's new theory is relevant as an answer to the question he had originally posed in appearance only. It must be noted that he lumps "deciding" and "deliberating" together in this instance as a single item. Now, deliberation may reveal to me that certain desires if pursued might be disharmonious with the universe of my interests and might modify the intensity of these desires, but *to deliberate is not to choose.* Deliberation only reviews the possibilities; it does not select among them. It does modify the outcome of the final "resolution" just as Ross suggests; but it is not the same thing as that resolution. Because deliberation does modify the outcome of the process, it can be agreed that determination of the choice by the strongest set of motives does not make *deliberation* "completely nugatory." But what about *choice*? If it is proposed that choice is determined by the strongest *set* of motives, this does not answer the original question, which was: "If action is described as following upon the strongest desire, is not resolution or choosing or deciding made completely nugatory? Would not the same act follow if the step called choosing were entirely omitted?" The question can still be asked of the new theory which Ross substitutes. If choice is determined by the strongest *set* of motives, is not resolution, choosing, deciding made completely nugatory? Would not the same act follow if the step called choosing were entirely omitted? Is not the outcome already decided before the choice takes place? Is it not decided at the moment when a given set of motives becomes dominant? It seems that Ross's new theory suffers from exactly the same difficulty which made him dissatisfied with the old theory! On *both*, the strongest set of motives should be sufficient to bring about action of themselves without the intervention of an act of choice if Ross and the determinists are correct!

V. Does Choice Precede the Appearance of the "Strongest Motive" or Set?

Although libertarianism has its serious difficulties, it at least does not suffer from the difficulty just discussed. It can explain why an act of choice can and must occur. Again the importance of determining which comes first, the choice, or the strongest motive or set of motives, must be emphasized. Although this question has not been stressed by previous libertarians, it is of crucial import.

Now, we must recognize that we never choose an end without having a motive for doing so. Libertarians have been criticized much for offering the oddity of "motiveless choice" to the philosophical world. Although some may have offered the theory, it is not here being offered. We never attend to any end as a live option for our choosing unless there is an interest in or desire for that end. If it is admitted that we never choose an end without *some* desire for that end, are we also admitting that we choose a given end *because* our desire for that end or the set to which it belongs, is the *strongest* desire or set of desires that we have? Not at all!

---

[11] C. D. Broad, *Five Types of Ethical Theory* (London, Routledge & Kegan Paul Ltd., 1952), p. 228.

[12] Ross, *op. cit.*, p. 228.

[13] *Ibid.*

[14] *Ibid.*

The libertarian admits that when our motives are clearly strongest, *then* choosing is superfluous. The libertarian has not always argued this way, but this is what he ought to maintain. Ross himself recognizes a species of action which follows directly upon desire without the intervention of choice.[15] Introspection reveals that when we know very clearly that we most strongly want a certain end, we do not have to make a choice in the matter at all. We simply pursue that end. It is only in cases where desires are strongly conflicting that we ever have to make a choice among contemplated alternatives, and the fact that in such cases desires have to be supplemented by choices indicates that no particular desire or set of desires is clearly strong enough until *after* the choice is made. We seldom have to make choices, for we usually know well what we want and pursue the wanted end without having to choose it at all. Choices are made only in cases of conflict, and the job of choice is to supplement a weak desire or set of desires and make it strongest over its competitors. When we choose, we *attend* closely to the alternative with which we identify ourselves and not to its competitors. This concentrated attention upon a certain end causes our desire for that end to increase in intensity, and lack of attention to the competing ends causes the desires for them to decline sharply and almost immediately in strength and often to pass almost completely out of existence. Thus *some* motive preceded choosing, but choosing preceded that moment when the motive which brings about the act becomes strongest. It makes sense to ask why a certain end is chosen, but the answer to this question can never be a complete causal explanation of the choice. We always choose an end because we desire that end. But there can be no explanation of why we chose it instead of its competitor which was equally desired. To insist that there has to be an additional motive for the choice would be to deny that the sets are equal, in which case no choice would occur. It is not choosing which is determined by the strongest motive, as determinists have traditionally argued. Rather, it is choosing which decides which motive is to become strongest. Determinism of the "strongest motive or set of motives causes choice" variety would make choosing superfluous, but in fact choosing definitely does have a job to do when it occurs.

A brief contrast between the theory of the relation of choice and motivation developed in the preceding paragraph and the theory as it is developed by C. A. Campbell will help me to explain my position more clearly. I shall discuss only Campbell's theory of the relation of choice and motivation in cases of non-moral choosing, though I strongly disagree with his theory of moral choosing also. Campbell tells us that "in all those practical choices—and they comprise perhaps 99 per cent of the choices in most men's lives—in which there is *no* felt conflict of duty with desire, it seems clear that the determinant of choice can only be the agent's strongest desire."[16] Campbell accepts all too uncritically the "determined by the strongest motive" pattern of explanation which is typical of the determinist. Even Ross appears to be much more critical of this pattern than is Campbell, as we have seen. Campbell does not appear to be in the least sensitive to the difficulties with the "strongest motive"—"strongest force" analogy which has been here discussed.

The chief motivation behind Campbell's attempt to treat cases of non-moral choosing as determined by the strongest motive is his desire to provide an explanation of "continuity of character," by which is meant that relatively stable and constant set of interests, habits, dispositions, abilities, etc., which make the individual behave in reliable, predictable ways; but a much more satisfactory account of this can be provided which will not involve us in the unhappy paradoxes of his explanation. Campbell explains that

> . . . self-activity of moral decision . . . differs very significantly from the self-activity of ordinary choices in virtue of the fact that while in both cases it is the self that is active, in the former case it is not the self merely *qua* formed character that acts, but the self as somehow *transcending* its own formed character.[17]

He says that his theory of freedom "leaves character still an enormously important factor in the moral life."[18]

There is really no need to go to such extremes to explain continuity of character. A *much* more plausible explanation, which introspection confirms, is that we simply act most of the time directly on the basis of our motives—without making choices at all. We have to make choices only in cases of serious conflict among motives or sets of motives, and such conflict is the exception rather than the rule. It is not 99 per cent of our *choices* which are determined directly by the strongest motives. It is rather 99 per cent of our *actions* which are so determined, *without our making choices at all.* Of course, Campbell's statistics may be a bit too high. Perhaps it is only 80 or 90 per cent of the time that we act in this "impulsive" sense. It is in any event choosing itself which is such a rare thing, though perhaps there are other modes of self-activity such as attending or setting ourselves to act which are omnipresent in waking life. It is choosing which introduces the greatest *dis*continuities of character of all the modes; and our character is continuous because we seldom have to perform the highly complex activity of choosing, not because most of our choosing is determined by the strongest motive or set. When choice does occur, even if it be in non-moral situations, there is the same transcending of character which Campbell has recognized in moral decision. It is only because character alone is not sufficient to do the job at hand and needs to be supplemented that we ever have to make decisions at all. Campbell admits that decision in non-moral instances as he interprets it is merely "formal rather than real."[19] It is here maintained in contrast that in those cases where choosing *would* be superfluous, it simply does not occur at all. Where inclinations are clearly stronger than their competitors, there is no conflict and no merely "formal" choice.

---

[15] *Ibid.*, pp. 192, 193.

[16] Campbell, *op. cit.*, p. 151.

[17] *Ibid.*, p. 150.

[18] *Ibid.*

An elementary mistake in analyzing the relationship between motivation and choice lies at the basis of other arguments which Campbell employs to support his theory that non-moral choices are determined by the strongest motive. He holds that the *only* possible function of choosing other than that of putting a merely formal cap on the strongest desire is that of acting *against* the strongest desire. Non-moral choosing does the former; moral choosing does the latter. Campbell asks:

> What possible motive *could* there be for a man to choose something different from that to which his desiring nature most strongly inclines him, except the fact that he deems this most strongly desired end to be somehow incompatible with his duty?[20]

Against Campbell, I would hold that we never choose against our strongest desire. The function of choosing is that of *supplementing* the strength of desires which would not be strong enough alone to predominate over competitors. Of course, we never choose something which we do not desire *at all*, and to this extent it may be agreed with Campbell that there is no choice where there is no inclination. But it cannot be agreed that non-moral choosing has no job to do whatsoever since it does not act *against* the strongest motive or set of motives. The function of non-moral choosing is not to act against the strongest motive, but this does not mean that it has *no* function. There is another interpretation. Its usual function is to "add its weight," to use a metaphor from physics, to the side of that desire for an anticipated goal which is not "heavy" enough of itself to outweigh its competitors. It is only because he overlooks this real function of non-moral choosing that Campbell can hold that it has no function whatsoever except that of putting a purely "formal" cap on the strongest set of inclinations.

University of Tennessee

Received October 19, 1965

---

[19] *Ibid.*, pp. 155, 156.

[20] *Ibid.*, p. 151.

American Philosophical Quarterly
Volume 4, Number 1, January 1967

IX. NATURAL RIGHTS

WILLIAM J. WAINWRIGHT

I

It may not be useful to distinguish natural rights from other rights. If one does choose to speak of natural rights (and many do), it is desirable to be as clear about them as possible. I propose to define natural rights as rights which meet three conditions. They must correspond to moral obligations, and these obligations must be both universal and important. Each of these conditions calls for some comment.

(A) Saying that natural rights correspond to obligations does not commit me to the view that wherever we have a right we have a corresponding obligation, but only to the view that wherever we have a natural right we have a corresponding obligation. Nor am I maintaining that wherever we have an important and universal moral obligation we have a natural right, but only that wherever we have an important and universal moral obligation, then, if there is a corresponding right, that right is a natural right.

(B) The moral obligations which correspond to natural rights are universal. By definition any moral obligation is universal in the sense that it is *prima facie* binding on anyone who is in the appropriate situation. Thus the obligation to keep promises is *prima facie* binding on anyone who has freely made a promise. Statements of moral obligation are universal or universalizable. We are not, however, interested in this sort of universality.

Someone who occupies a certain station in life may have moral obligations to those who stand in some special relation to him. Obligations of this sort are universal in the sense that anyone who occupies that station has certain definite obligations to anyone who stands in the special relation to him. But there are two senses in which these obligations are not universal. The obligations do not in fact fall on every man (or they fall on them only conditionally—*if* they ever occupy that station then they must . . .) and one has these obligations only toward certain persons (those who stand in the special relation). The demand that an obligation be universal may be, and I think sometimes is, a demand that it either fall on all men actually as well as conditionally (that it be universal with respect to its subject) or that it be an obligation of which all men are the object (that it be universal with respect to its object) or both.

1. What can it mean to say that an obligation actually falls on all men? It can mean at least two things: (a) That the obligation falls on man as man and not simply in so far as he plays a special role or enters into a special relation with someone, or (b) that the role or special relation which gives rise to the obligation is one which most men are likely to assume at one time or another, for example, by being a parent or entering into an agreement.

2. That an obligation is one of which all men are the object can also be understood in two ways. That is, (a) it can be understood as an obligation which we have toward a man simply because he is a man and not because, for example, he has merit or has entered into a contract with us. (The obligation to save a drowning man is, I suppose, universal in this sense.) Or (b) it can be understood as an obligation which is such that most men would at some time or other be the object of this obligation, for example, by being a child or the recipient of a promise.

We should notice that if an obligation is universal with respect both to its subject and its object, it is reasonable to suppose that it is one which actually and not simply conditionally applies in all societies (and this in the double sense that both subjects and objects of the obligation are to be met with in all societies).

If the obligation is one which falls on man as such, then it is an obligation which falls on men in all societies. If the obligation is universal with respect to its subject in the second sense, i.e., if it is an obligation which most men fall under at some time or other, then it is natural to expect that the obligation will fall on men in all societies. However, this is not necessarily the case, for it could be true of a given obligation both that it falls on most men at some time or other and that in some society it never falls on anyone. This observation suggests another sense in which an obligation could be universal with respect to its subject—viz., that it be an obligation which falls on most men at some time or other in *any* society. (Parental obligations or those arising from a promise *might* be obligations of this sort.) This in turn could be taken in at least two ways. "Any society" could be understood as meaning any actual society (past societies, present societies, and those which are likely to arise in the future) or it could be understood as meaning any possible society. Thus we have four senses in which any moral obligation is universal with respect to its subject. The four senses are:

i. that the obligation fall on man as man,
ii. that it fall on most men at some time or other,
iii. that it fall on most men at some time or other, in any actual society, and
iv. that it fall on most men at some time or other in any possible society.

Corresponding to the four senses in which an obligation may be universal with respect to its subject, we can distinguish four senses in which an obligation may be universal with respect to its object:

i. the obligation is to man as man,
ii. it is such that most men would at some time or other be the object of this obligation,
iii. it is such that in any actual society most men will at some time be its object, and
iv. it is such that in any possible society most men will at some time be its object.[1]

I have said that natural rights correspond to universal obligations. We are now in a position to spell this out more clearly.

Natural rights may be provisionally defined as rights which correspond to important moral obligations which are universal with respect to their object in either the first or fourth senses and which are universal with respect to their subject in either the first or fourth senses.

(C) By definition any moral obligation is important in the sense that it overrides (all?) non-moral obligations. Natural rights not only correspond to moral obligations, and thus share in the importance which is involved in any moral obligation, they correspond to the more important of these obligations. One obligation is more important than another if it generally overrides the other where there is a conflict between the two.

The universality of the corresponding obligation might be thought to be sufficient to distinguish natural rights from other rights, particularly in view of the fact that even the less important obligations are important in the sense that they override all non-moral obligations. Importance has been made a condition for the following reason. I am inclined to think that "Be polite" is a moral rule. It is possible that the ground for having a code of etiquette, and this rule, would be found in all possible societies. The obligation would then be universal in the required sense. Furthermore, there seems to be nothing improper in speaking of a right to be treated politely (e.g., not to be insulted). A right, therefore, corresponds to the universal obligation. Nevertheless it seems a bit odd to speak of a natural right to be treated politely. If we introduce the notion of importance, we get out of this difficulty, for presumably politeness is not one of the more important obligations.

I would like to conclude this section of my paper by comparing my definition with two others with which it might be confused. The first is the notion that natural rights are those which must be respected if society is to endure. For an obligation to have subjects and objects in a possible society, all that is necessary is that the grounds occur in that society. It is not necessary that the obligation be respected or even recognized. Thus, even if there were a society which did not recognize the right to life, it could still be the case that the obligation has subjects and objects in that society and that it is important in that society, i.e., it could be true both that men in that society ought not to kill and that men in that society have a right to life, and that this obligation and right should be regarded as overriding in that society. From this it follows that an obligation can be important and universal in the required way (and thus that the corresponding right is a natural right) and yet not be recognized in some society. Now if a right is not recognized in some society, then respect for that right can hardly be a necessary condition of the survival of society. Therefore, if my definition of natural rights is correct, their recognition is not necessary for the survival of society. Furthermore, I do not think that the claim that natural rights are those which must be respected if society is to endure is at all plausible. Respect for a right cannot be a necessary condition of society if it is not recognized and more or less adhered to in all societies. To accept the claim would mean that nothing would count as a natural right which was not recognized in all societies and I do not believe that we wish to place this restriction on the notion of natural rights.

---

[1] In saying that an obligation to do *x* is subject or object universal, I mean that (all or most) men are subjects or objects of a *prima facie* obligation to do *x*. I do not mean to deny that the obligation may be overridden by other obligations in certain circumstances.

Margaret MacDonald suggests[2] that natural rights are the fundamental conditions of a good society. If taken literally, this is probably false. The fundamental conditions for a good society may well include many things other than the recognition and fulfillment of certain obligations. For example, certain material and historical conditions may be necessary for a good society. On the other hand, since natural rights correspond to important moral obligations which occur in any possible society, and since no society can—perhaps analytically—be a good one which does not respect the important obligations which occur in that society, it follows that no society can be a good society which does not respect natural rights. Respect for natural rights is a necessary condition for a good society.

II

I shall now examine some of the rights which Locke considers to be natural rights. The purpose of this discussion will be to sharpen and qualify the remarks made in Part I.

(A) The right to life and the obligation not to take a human life rest on the claim that each of us is God's property.[3] If this is true at all, then it is true of each man in any possible society and the obligation will presumably be universal with respect both to its subject and its object in the first sense.

Liberty is also a natural right. To enslave another is to exercise the power of life and death over him. But because I have no right to take his life (he is God's property), I cannot have a right to enslave him.[4] Since the obligation not to kill another is universal with respect both to its subject and to its object in the first sense, it would seem that the obligation not to enslave another is also subject universal and object universal in the first sense.

But this will not do, for Locke says that men may forfeit their right to life and liberty if they unjustly attack another,[5] thereby cutting themselves off from the moral community.[6] This suggests that a necessary condition of the possession of the rights to life and liberty is membership in a community of rational beings who are governed by laws, either those of nature or those of political society.[7] If so, the ground for the right to life and liberty is not simply that one is God's property but also includes the fact that one is a member of a moral community in the sense required.[8] We have already seen that the first part of the ground will be found in all possible societies, but would the second part of the ground also be found in all possible societies?

A society most of whose members prey upon other communities is surely a possible society. In such a society most members would have unjustly resorted to violence and so would have forfeited (not have) the right to life and liberty. If my definition is correct, it would follow that life and liberty are not natural rights, for the corresponding obligations would not be object universal in either the first or fourth sense. We might conclude from this that my definition is inadequate. On the other hand, we might conclude that, contrary to Locke, refusing to resort to the unjust use of violence is not a ground for these rights. Or we might conclude that there is no natural right to life and liberty as such, but only a natural right to life and liberty in so far as one does not unjustly resort to violence. Members of a criminal society may well enjoy these rights.

(B) The primary ground for property rights is labor.[9] Locke also suggests gift,[10] bequest,[11] and exchange[12] as grounds for property rights. The obligations to respect property so acquired seem to be subject universal in the first sense. All men are obligated to respect property, i.e., one doesn't have to take up a special role or do something special or belong to a special society to come under this obligation. If gift, bequest, and exchange are sources of property rights, and labor is construed as broadly as Locke construes it (so as to include picking something up, gathering berries, etc.) then it is difficult to believe that there would be very many who would not have property in something in any possible society. If so, obligation will be object universal in the fourth sense.

---

[2] Margaret MacDonald, "Natural Rights" in *Philosophy, Politics and Society*, ed. Peter Laslett (First Series, Oxford, Basil Blackwell, 1963), pp. 47–48.

[3] John Locke, *Second Treatise on Civil Government*, para. 6.

[4] *Ibid.*, paras. 22–24.

[5] *Ibid.*, paras. 16–18, 23, 85.

[6] *Ibid.*, paras. 8, 10–11, 16.

[7] Cf. *ibid.*, para. 57.

[8] This is not quite right, for infants and madmen are not members of the moral community in any full sense (paras. 57, 60) and yet they have a right to life. They have no right to liberty—to act autonomously—but presumably they do have a right not to be enslaved. The difference is perhaps this: children and madmen are potentially full members of the moral community, but those who unjustly resort to violence have wilfully severed their connections with it.

[9] Locke, *op. cit.*, paras. 27 ff.

[10] *Ibid.*, para. 46.

[11] *Ibid.*, para. 72.

(C) Parents have a right to exercise authority over their children during the minority of the child.[13] What obligation corresponds to this right? The right arises from the obligation to care for one's children, but this is not a corresponding obligation in the sense in question (where rights in one person correspond to obligations in another). If we take seriously the notion that one is under the natural law only if one has reason, and if children lack reason,[14] then the corresponding obligation is not the obligation to obey one's parents. The corresponding obligation will presumably be that of not interfering with the parents' direction of their children. The ground for this right and obligation is the obligation parents have to care for their children. This in turn is based on the fact that children lack reason and have been given by God to the parents as their responsibility.[15] (Freedom is justified only in so far as one is a full member of a community governed by laws, but full membership requires rationality which children lack.[16] Furthermore, since children lack reason, guidance is necessary for their own good.[17]) Would these grounds be found in any possible society?

There are two difficulties. In the first place, a society in which parents do not know their own children is a logically possible society. For example, the children are taken away at birth and raised in public institutions. No birth records are kept. In the second place, it is logically possible that there be a (human) society in which children have no parents. (They are raised in a test tube.) In either case, the obligation which corresponds to the right may still be subject universal in the first sense. All men may be obligated to refrain from interfering with a parent's direction of his children. On the other hand, in the first society parents do not direct their children and are unable to do so, except accidentally, because they do not know who their children are. It would, therefore, seem that the obligation has *no* objects in this society and is therefore not object universal in either the first or the fourth sense. It should be noticed, however, that the reason why the obligation has no objects in this society is to be found in the peculiar institutions of this society, and these institutions may be morally suspect. It is possible, though perhaps false, that wherever the grounds for this right and obligation do not occur, the absence of these grounds is itself a ground for moral criticism. In the second society there are no parents and hence no objects of the obligation in question. The obligation is not, then, object universal in either the first or fourth sense. But it should be noticed that it may still be true that the ground for the right and obligation will be found in any society which is at all likely to occur. (The notion of a society which is at all likely to occur is to be distinguished from that of an actual society mentioned in Part I. Actual societies include past and present societies, and those which will probably arise in the future. Societies which are at all likely to arise include not only actual societies, but also all those societies the occurrence of which is not highly improbable. The dividing line between the two conceptions is clearly not a sharp one.)

(D) Locke also asserts that children have an obligation to honor and assist their parents on the grounds that they have been begotten and nurtured by them.[18] Corresponding to this obligation, parents have a right to these things.[19] I am inclined to think that Locke would say that the obligation is object universal in the fourth sense. It arises from the fact that parents have begotten and nurtured children and this is something most, though not all, men will do in all possible societies. Whether this is true or not would depend on whether the ground is primarily the begetting or the nurture of children. If it is the former, then the ground of the obligation will occur in any society which is at all likely to occur (though not in all possible ones, since societies without parents are logically possible). If it is the latter, then it need not occur in any society which is at all likely to occur—whatever the concept of a society which is at all likely to occur includes, it surely includes a society where the children are nurtured by the community. But even though the obligation to honor and assist one's parents might not be object universal, either in the sense that it has objects in all possible societies or in the sense that it has objects in any society which is at all likely to occur, the individual's obligation to honor and respect the ones who have nurtured him may be, for presumably in any possible society there would be objects of this obligation.[20] At the same time, there could well be a society in which a relatively few people were involved in nurturing the children, and if this were the case it would not be true that *most* people in all possible societies were the object of this obligation.

---

[12] *Ibid.*, paras. 14, 46.

[13] *Ibid.*, para. 67.

[14] *Ibid.*, para. 57.

[15] *Ibid.*, para. 56.

[16] *Ibid.*, para. 57.

[17] *Ibid.*, para. 63.

[18] *Ibid.*, para. 66.

[19] *Ibid.*, para. 67.

If the primary ground for the obligation is that one has been begotten by one's parents, the right will not be subject universal in either the first sense or the fourth sense because of the possibility of a parentless society. However, all men will be under it in any society which is at all likely to occur. If the ground for the obligation is nurture, then it need not be subject universal in either the first or fourth sense nor in the sense that persons would be subject to this obligation in any society which is at all likely to occur, for there may be societies in which parents do not nurture their children. On the other hand, an obligation to honor and assist those who have nurtured you will be subject universal in the first sense (if we may suppose that all men must be nurtured to a greater or less degree by someone or other).

(E) Our discussion of Locke has suggested that there may be senses of universality other than those which were suggested in Part I. These senses are:

v. the obligation has (many) subjects and objects in any society which is at all likely to occur,
vi. the obligation has (only) some subjects and objects in any possible society,
vii. the obligation has (only) some subjects and objects in any society which is at all likely to occur, and
viii. any society in which the obligation does not have (many or some) subjects or objects is such that the features which constitute reasons why there are no subjects and objects are grounds for moral criticism.

Should we abandon our provisional definition and extend the notion of natural rights so as to include rights which correspond to important obligations which are universal in any or all of these senses? I am inclined not to do so, on the grounds that natural right theorists have had a very strong sort of universality in mind and that the provisional definition is therefore closer to what has traditionally been meant by "natural rights." However, it seems to me to make little difference whether or not we extend the notion of natural rights in the ways suggested, as long as we are clear as to just what we are doing.

(F) Are the obligations which correspond to the rights we have discussed important obligations? The obligations not to kill, enslave or steal from another are surely important. The obligation to assist, honor, and respect one's parents and the obligation not to interfere with a parent's direction of his children seem less important. In particular the last obligation seems to be one which could be easily overriden by other more important obligations. I am not sure whether it is more plausible to deny that the last is a natural right or to drop the notion that a natural right must correspond to a more important obligation.[21]

III

I should now like to say something about the connection between the notion of natural rights and the notion of a state of nature. The concept of a state of nature can be used in at least two ways. It can be used to argue for the necessity of political society by pointing out the disadvantages which men would suffer if there were no political society. The notion of a state of nature can also be used in another way, for it may be argued that by examining man in a state of nature we can discover natural rights and obligations in the defined sense. Locke uses the concept of a state of nature in both these ways.

---

[20] One might object that the nurture of children is not a *logically* necessary condition of a human society. I am inclined to think that childhood (involving nurture and guidance of some sort) is (logically) necessary for a human existence in a way in which begetting children and being begotten by human parents is not. I am, however, by no means clear on this point.

[21] We might have considered (putative) natural rights other than those mentioned by Locke. We might, for example, have considered the rights mentioned in the U.N. Declaration. I think that an examination of these rights would show that they correspond to important obligations which are subject and object universal in sense (i) or in one of the senses (iv) through (viii). It is difficult, however, to determine in just what senses a corresponding obligation is universal without determining what the grounds of the obligation are. In examining Locke, one has the advantage that the grounds are given.

If we look at our characterization of natural rights we can see the plausibility of the second move. Natural rights correspond to obligations which fall on men in any society and whose objects can be found in any society. The picture of a state of nature can (in part) be regarded as an attempt to sketch features common to all societies. (Examples of such features might be parental relationships, work, agreements entered into, etc.) If these features are the grounds of certain rights and obligations, one may conclude that these rights and obligations are universal in the required sense, that they will hold in any society. Furthermore, if these rights and obligations are overriding in the state of nature, we may expect them to be overriding in all societies. This is not, however, strictly necessary, for the special arrangements of some society could be such that the obligation or right arising from one of the common features was generally outweighed in that society though not in others.

For the purpose of determining natural rights in the defined sense it is not essential that the state of nature be a state without government. The picture will be drawn without government only if political society is considered to be adventitious, i.e., if political society is not considered necessary to human society. That the state of nature has in practice been pictured as a state without government is due both to the fact that political society was considered accidental or adventitious and to the fact that the examination of a state of nature was serving a double purpose—furnishing reasons for the necessity of government as well as a doctrine of natural rights.

When the state of nature is pictured as society with government abstracted, we have another meaning of "natural rights," viz., "natural rights" are those moral rights, perhaps the most important ones, which obtain in a state of nature whether or not these rights carry over into, or remain important in, political society. (The right to judge one's own case would be an example of a right which one has in the absence of judicial institutions and which does not carry over into political society.) It is to be noticed, however, that if a right does not carry over into political society or if it loses its importance in political society, then it is not a natural right in the defined and, I venture to say, important sense.

It is true that Locke is interested in drawing a distinction between those rights and obligations which arise in nonpolitical society and those which arise in political society. It is not clear that we should be interested in this. In the first place, it may well be that any possible society is a political society. (This of course does not imply that certain rights may not be grounded in the nonpolitical aspects of society.) In the second place, Locke is interested in natural rights largely because he is interested in showing that there are (should be) restraints on arbitary power. Yet the only rights which will function in this way are those which carry over into political society and are thus natural rights in my sense. Finally, natural rights are usually thought to be rights which all or most men have. If this is the case, rights which hold only in a state of nature cannot be natural rights.

University of Illinois

Received January 4, 1966

AMERICAN PHILOSOPHICAL QUARTERLY

Edited by

NICHOLAS RESCHER

With the advice and assistance of the Board of Editorial Consultants:

William Alston
Alan R. Anderson
Kurt Baier
Lewis W. Beck
Nuel D. Belnap, Jr.
Roderick M. Chisholm
L. Jonathan Cohen
James Collins
James M. Edie
Peter Thomas Geach
Adolf Grünbaum
Carl G. Hempel
John Hospers
Raymond Klibansky
Ernan McMullin
Benson Mates
John A. Passmore
Günther Patzig
Richard H. Popkin
Wesley C. Salmon
George A. Schrader
Wilfrid Sellars
Alexander Sesonske
J. J. C. Smart
Manley H. Thompson, Jr.
James F. Thomson
G. H. von Wright
John W. Yolton

VOLUME 4/NUMBER 2 APRIL 1967

CONTENTS

I. HECTOR-NERI CASTANEDA: *Indicators and Quasi-Indicators* 85
II. NORWOOD R. HANSON: *The Genetic Fallacy Revisited.* 101
III. DEWEY J. HOITENGA, JR.: *Logic and the Problem of Evil* 114
IV. TIMOTHY DUGGAN AND BERNARD GERT: *Voluntary Abilities* . . . 127
V. G. H. VON WRIGHT: *Deontic Logics* . 136
VI. D. C. EMMONS: *Justice Reassessed* . 144
VII. GEORGE SCHLESINGER: *The Probability of the Simple Hypothesis* . 152
VIII. NELSON PIKE: *Hume's Bundle Theory of the Self: A Limited Defense* . 159
IX. GARETH B. MATTHEWS: *The Inner Man* 166
X. D. J. O'CONNOR: *Tests for Intentionality* 173

PUBLISHED BY BASIL BLACKWELL WITH THE COOPERATION OF THE UNIVERSITY OF PITTSBURGH

AMERICAN PHILOSOPHICAL QUARTERLY

POLICY

The *American Philosophical Quarterly* welcomes articles by philosophers of any country on any aspect of philosophy, substantive or historical. However, only self-sufficient articles will be published, and not news items, book reviews, critical notices, or "discussion notes."

MANUSCRIPTS

Contributions may be as short as 2,000 words or as long as 25,000. All manuscripts should be typewritten with wide margins, and at least double spacing between the lines. Footnotes should be used sparingly and should be numbered consecutively. They should also be typed with wide margins and double spacing. The original copy, not a carbon, should be submitted; authors should always retain at least one copy of their articles.

COMMUNICATIONS

Articles for publication, and all other editorial communications and enquiries, should be addressed to: The Editor, *American Philosophical Quarterly*, Department of Philosophy, University of Pittsburgh, Pittsburgh, Pennsylvania 15213.

OFFPRINTS

Authors will receive gratis two copies of the issue containing their contribution. Offprints can be purchased at low cost through arrangements made when checking proof.

SUBSCRIPTIONS

The price *per annum* is six dollars for individual subscribers and ten dollars for institutions. Checks and money orders should be made payable to the *American Philosophical Quarterly*. Back issues are sold at the rate of two dollars to individuals, and three dollars to institutions. All correspondence regarding subscriptions and back orders should be addressed directly to the publisher (Basil Blackwell, Broad Street, Oxford, England).

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 2, April 1967

I. INDICATORS AND QUASI-INDICATORS[1]

HECTOR-NERI CASTANEDA

1. *Purpose, problem, and conventions.* Crucial to many problems in the philosophy of mind, epistemology, and metaphysics are: (A) a clear understanding of the roles and interrelations of the types of words we employ to refer to the entities we encounter, and (B) a clear understanding of the logic of cognitive and linguistic verbs (like 'think,' 'suppose,' and 'infer,' on the one hand, and 'say,' 'inform,' 'argue,' and 'show,' on the other). My ultimate purpose here is to make a contribution toward both (A) and (B).[2]

My contribution toward (A) consists mainly in showing, via principles (Q.7)–(Q.10) below, that a certain mechanism of reference to particulars, which is indispensable to a conceptual scheme that allows for the possibility of other persons or selves, is a unique logical category of reference not reducible to the better-studied mechanisms of reference, namely, demonstrative pronouns or adverbs, proper names, singular descriptions, and ordinary variables of quantification. This irreducible mechanism is constituted by what I call quasi-indicators. Following Goodman,[3] I call *indicators* the personal and demonstrative pronouns and adverbs like 'this,' 'that,' 'I,' 'you,' 'here,' 'there,' and 'now,' when they are used to make a strictly demonstrative reference, i.e., when they are used purely referentially either to single out an item present in the speaker's current experience or to pinpoint a self that is a relatum in the cognitive relation evinced by the speaker. I call *quasi-indicators* the expressions which in *oratio obliqua* represent uses, perhaps only implicit, of indicators, i.e., uses which are ascribed to some person or persons by means of a cognitive or linguistic verb. For example, suppose that at a certain place p and time t, A says to B: (1) "I am going to kill you here now." Suppose that C reports this by asserting: (2) "A said to B at p at t that he was going to kill him there then." Sentence (2) contains the quasi-indicators 'he,' 'him,' 'there,' and 'then,' which represent uses, ascribed to A, of the indicators 'I,' 'you,' 'here,' and 'now,' respectively. We shall see that the very same marks or noises 'I,' 'you,' 'here,' 'there,' 'he,' 'this,' function as either indicators, quasi-indicators, or something else. It is a mere accident of grammar that the same physical objects are used in different logical roles. The underlying rationale is this: indicators are a primary means of referring to particulars, but the references made with them are personal and ephemeral; quasi-indicators are the derivative means of making an indexical reference both interpersonal and enduring, yet preserving it intact.

My contribution toward both (A) and (B) together consists of the formulation, with as much precision as I can here, of the fundamental principles of the logic of both indicators and quasi-indicators in *oratio obliqua.* In this regard principles (I.5), (I.12)–(I.14), (Q.2), (Q.11), and (Q.12) are the most important of the lot.

I have aimed at an exhaustive treatment of the basic logic of singular indicators and quasi-indicators in *oratio obliqua*, except for the treatment

---

[1] Part of this research was carried out under the National Science Foundation Grant No. GS–828.

[2] Some of the main results of this paper are generalizations for all quasi-indicators of previous results obtained, in the terminology of (Q.4) below, for the very special primary quasi-indicator 'he*.' These previous results together with other complementary results are discussed in my " 'He': A Study in the Logic of Self-consciousness," forthcoming in *Ratio.* It will be cited as "*He.*"

I owe the initial stimulus for these generalizations to my colleague Robert Sleigh, Jr., who on March 12, 1965, said to me that 'then' in "Jones believed at t that it was raining then" is like (the primary quasi-indicator) 'he*' in not being replaceable by its antecedent 't.' Peter T. Geach has really studied *He* and in correspondence has provided a sustained flow of encouragement, criticisms, and suggestions. George Nakhnikian read the first draft of this paper and pointed out several obscurities and errors. Wilfrid Sellars read the second draft and pointed out errors and obscurities and suggested improvements.

[3] N. Goodman, *The Structure of Appearance* (Cambridge, Mass., Harvard University Press, 1951), p. 290. However, Goodman's criterion for the selection is different: "a word is an *indicator* if (but, . . ., not necessarily only if) it names something not named by some replica of the word."

of the obvious indicators and quasi-indicators included in verbal tenses.[4]

To simplify the discussion I shall abide by the following conventions: (1) single quotes will be used to form names of sentences or expressions; (2) double quotes will be used to form names of statements made or that can be made by uttering the quoted sentences on given occasions; (3) numerals within parentheses will be used as either names of sentences or names of statements, the context indicating which one is meant; (4) the phrase 'statement of the form "..." ' will be short for 'statement made by the sentence of the form "..." (if assertively used by a certain person on a certain occasion)'; (5) an asterisk ' * ' after a word or phrase indicates that the word or phrase functions in a given sentence as a quasi-indicator; (6) 'statement' and 'proposition' will be used interchangeably in the sense these words have been used by Strawson and others.[5]

2. *Indicators*. Let us start our investigation by listing some well-known properties of indicators. These properties determine the special roles of indicators in *oratio obliqua*.

(I.1) Indicators formulate a personal reference to a single entity made primarily by the speaker, made derivatively by the hearers, and not made at all by other persons. Here we shall be concerned with primary reference only.

(I.2) Reference to an entity by means of an indicator is purely referential, i.e., it is a reference that attributes no property to the entity in question. Since an indicator expresses an act of placing an item in a person's experience or expresses an act of pinpointing the person who enters in a current cognitive relation, it can at most be said that the use of an indicator indirectly, or implicitly, attributes to an entity the property of being a thinker or the property of being placed in a thinker's (or speaker's) experience at the time of thinking or speaking.

(I.3) Reference by means of an indicator is neither identical with, nor equivalent to, reference by means of descriptions that contain no indicators. By (I.2) reference by means of an indicator could at most be equivalent to reference by means of a description that has no indicators but attributes to the referred entity just the property of being a thinker or of being placed in a thinker's (or speaker's) experience at the time of thinking (or speaking). But this is, of course, false. On the one hand, this attribution by being direct goes beyond what a mere indicator can accomplish. On the other hand, such a description would have to have a word that refers to the thinker (or speaker) in question without in the least attributing to him any property at all, and similarly for the time of thinking. But since the description under consideration has no indicators, there is no way for it merely to refer to the thinker or to the time of his thinking. Thus, at least one sentence *S* containing an indicator *I* formulates or can formulate, as uttered on a certain occasion, a statement not equivalent to the statement formulated by any other sentence that differs from *S* by having descriptions with no indicators instead of *I*.

(I.4) Indicators have a *referential priority* over all names and descriptions. A person may use a name or description correctly, and yet fail to refer to the object to which he purports to refer because there may be not one, but many, objects which have the name in question, or the properties mentioned in the description. However, a man who uses an indicator correctly cannot, because of a multiplicity of candidates for reference, fail to refer to the object he purports to refer to.

(I.5) The first-person pronoun has a strong *ontological priority* over all names, contingent descriptions, and other indicators. A correct use of 'I' cannot fail to refer to the entity it purports to refer; moreover, a correct use of 'I' cannot fail to pick up the category of the entity to which it is meant to refer. The first-person pronoun, without predicating selfhood, purports to pick out a self *qua* self, and when it is correctly tendered it invariably succeeds. On the other hand, correctly used names, contingent descriptions, and non-first-person indicators may fail either to even pick out a referent or to pick out the intended category. The time- and space-indicators, e.g., 'now' and 'here,' have a weak ontological priority over names, descriptions, and other categories of third-person indicator. They succeed in always picking out a time or a place, although they may in principle fail to pick out the category of physical or external time and space as in a dream. (These ontological priorities are one of the fundamental facts underlying Descartes' *Cogito* and Kant's theses on the transcendental self and on space and time as forms of perception.)

---

[4] The latter should offer no major difficulty after both the study of "now" and "then" included here and the study of tenses by W. Sellars in "Time and the World Order," in H. Feigl, M. Scriven, and G. Maxwell (eds.), *Minnesota Studies in the Philosophy of Science* (Minneapolis, University of Minnesota Press, 1957). An important problem open in this area is the examination of difficulties (if any) raised by quasi-indicators for Quine's program of eliminating all singular referring terms in favor of predicates and variables of quantification. See his *Word and Object* (Cambridge, Mass., Massachusetts Institute Technology Press, 1960), pp. 163, 170–186. The auxiliary notation introduced here can be easily formalized and, then, adjoined to a calculus of "knows" and "believes," such as Hintikka's in his *Knowledge and Belief* (Ithaca, New York, Cornell University Press, 1963). For the need to supplement Hintikka's calculus with some such notation, see my review of his book in *The Journal of Symbolic Logic*, vol. 24 (1964), pp. 132–134.

[5] Cf. P. F. Strawson, *Introduction to Logical Theory* (London, Methuen & Co., 1952), and cf. R. L. Cartwright, "Propositions," in R. J. Butler (ed.), *Analytic Philosophy* (Oxford, Blackwell, 1962).

It might be induced that this contrast between 'I' and the other indicators is misguided, on the ground that 'I' is analyzable in terms of 'this' or 'here now.' Reichenbach, for instance, claims that "the word 'I' means the same as 'the person who utters this token'."[6] This claim is, however, false. A statement formulated through a normal use of the sentence 'I am uttering nothing' is contingent: if a person utters this sentence he falsifies the corresponding statement, but surely the statement *might*, even in such a case, have been true. On the other hand, the statements formulated by 'The person uttering this token is uttering nothing' are self-contradictory: even if no one asserts them, they simply cannot be true.

The first-person indicator is unanalyzable. On the one hand, if 'ϕ' stands for an analytic predicate, i.e., a predicate such that the statement of the form "Everything is ϕ" is analytic or necessarily true, then, obviously, the first-person indicator 'I' cannot be defined as 'the only person who is ϕ.' On the other hand, if 'ϕ' stands for a contingent predicate, then the normal singular statements of the form "I am ϕ" assertable by human beings are contingent. This is so, because all such statements entail the corresponding singular statements of the form "I exist," which are contingent, since they can be false, i.e., since the corresponding singular statements of the form "I might have not existed" are true. Hence, the corresponding singular statements of the form "I am not-ϕ" are also contingent. Yet the corresponding statements of the form "The only person who is ϕ is not-ϕ" are self-contradictory. Hence, the first-person indicator cannot be analyzed in terms of any other expressions whatever; *a fortiori*, it cannot be analyzed in terms of the other indicators. In particular, 'I' cannot be analyzed as 'this self' or 'the self here now': a person's statements of the form "I am not-a-self" are contingent, while his statements of the form "This self (or the self here now) is not-a-self" are self-contradictory. (The unanalyzability of 'I', especially its unanalyzability in terms of [third-person] demonstratives, is another fundamental fact underlying the idea of the transcendental self.)

Evidently, parallel arguments hold for the second-person indicators, the third-person indicators referring to persons or objects, and the time- and space-indicators. For instance, the object-indicator 'this' cannot be defined in terms of any description, since the indicator attributes no property at all to the object it refers to; and it cannot be defined merely indexically as 'here now,' for statements of the form "This [meaning an object] is not here now" are contingent, while statements of the form "What is here now is not here now" are self-contradictory, if the two occurrences of 'here now' have the same reference.

It must be noted that the words 'now' and 'here' are sometimes replaced by 'this,' as e.g., in "Here (this) is a good place to park." But this in no way argues against the unanalyzability of the time- and space-indicators. What we have here are two interesting features of ordinary language: (i) the grammatical fact that 'here,' 'there,' and 'now' are primarily adverbs, and lack, therefore, the syntactical flexibility of substantival pronouns, so that in certain constructions they are replaceable by some other expression; (ii) the words 'this' and 'that' have two indexical uses: (a) as general signs of third-person indexical reference, and (b) as categorical indicators of third-person indexical reference to objects. It is in their use (a) that 'this' can replace 'now' and 'here,' and 'that' can replace 'there.' Obviously, 'here,' 'there,' and 'now' are not definable in terms of use (b) of 'this' or 'that.' For instance, 'here' is not analyzable as '(at) the place occupied by this object,' since "Nothing is here" is contingent, while "Nothing is at the place occupied by this object" is self-contradictory.

In sum,

(I.6) There are five irreducible indexical roles, whatever noises or marks we may employ to discharge them: first-person, second-person, third-person, (specious) present-time, and (speciously) presented place. [(I.6) is supplemented by (I.15) below, after (Q.5).]

---

[6] H. Reichenbach, *Elements of Symbolic Logic* (New York, Macmillan, 1947), p. 284.

(I.7) The pronoun 'I' and all descriptions primarily, and all names derivatively, have an *epistemological priority* over all other indicators. A person's own uses of the first-person pronoun differ radically from his uses of the other indicators. One's knowledge or belief about oneself must be in the first-person form for it to be really self-knowledge or self-belief. On the other hand, to retain knowledge or belief, or merely to re-think, of the objects or persons originally apprehended by means of indicators, one must reformulate one's knowledge or belief, or thought, of those objects in terms of names or descriptions, or in terms of 'I.' *The references made by an indicator other than 'I' are ephemeral, and necessarily eliminable for those who make them.* Of course, the elimination of an indicator may very well introduce another one, as, for instance, when a man thinks something of the form "This is blue," and later on re-thinks it as, e.g., "The object I saw *there* an hour ago was blue." Elimination is here a process of preserving information, not a process of analysis or of literal translation. By (I.6) we already know that there is no analysis of the indexical role.[7]

(I.8) A person's uses of indicators are first-personal and eliminable for every other person.

3. *oo-prefixes.* Now we shall discuss some general properties of *oratio obliqua* to which we shall continuously refer in the ensuing investigation.

Cognitive acts or attitudes or dispositions are second-order tetradic relations involving a subject (person or self), a place, a time, and a proposition. Linguistic acts are second-order relations involving typically five or more relata: a speaker, one or more addresses, a place, a time, and a proposition. (I call them all second-order relations because one of the relata is a proposition.)

In a sentence formulating a proposition to the effect that a cognitive or linguistic relation obtains, we shall distinguish two parts: (i) an *oratio obliqua*, and (ii) an *oratio obliqua* prefix. *Oratio obliqua* is the clause that formulates or represents the propositional relatum of the cognitive or linguistic relation. A *strict oratio obliqua prefix*, or simply *strict oo-prefix*, is a clause that contains: (a) a cognitive or linguistic verb, (b) expressions denoting the non-propositional relata, and (c) the word 'that' at the right-hand end. For example, in

(1) Grabbing the knife on the table Peter told Mary in the kitchen at 3:00 p.m. that he would kill her with it there,

the strict oo-prefix is 'Peter told Mary in the kitchen at 3:00 p.m. that.' In (1) 'Grabbing the knife on the table' is not part of the strict oo-prefix. Clearly, (1) is really the conjunction "Peter grabbed the knife on the table at 3:00 p.m. yesterday, and (he) said to Mary in the kitchen at 3:00 p.m. yesterday that he would kill her with it there." Now, a clause containing a strict oo-prefix and other expressions will be called an extended oo-prefix.

For the sake of brevity we shall sometimes represent an oo-prefix by the schema

$$L[x, y_1, \ldots, y_n, p, t]\ (\text{——}),$$

where x is the speaker or thinker (arguer, knower, etc.), $y_1, \ldots, y_n$ are the addressees, p the place, and t the time involved in the linguistic or cognitive relation. If the relation is cognitive, 'y_1', . . ., 'y_n' vanish. Here '——' indicates the place of the *oratio obliqua* involved.

Linguistic acts of the assertive kind, i.e., whose objects are propositions, not sentences or noises, are dependent on cognitive acts. For example, to state or say or claim is at once to have the thought that something is the case. It is clear that a noise or mark by itself is not an indicator: it is one if and only if it carries with itself an indexical reference. Likewise, a word is not a name or description unless it carries within itself a nominal or descriptive reference that some speaker means it to carry. Thus, the roles which words or phrases or clauses discharge as vehicles of components of a proposition depend on the structural relations of these components *qua* components of the proposition in question. The structure of what is thought, the proposition, is the fundamental structure, and in principle one could think without using a language as a means of thinking.[8] Nevertheless, by the same token, even if one can think without using a language at all, the propositions one deals with in mere non-linguistic thinking have a structure which parallel certain structures of the complete sentences in which one would formulate one's thoughts, if one had a language.[9] We may suppose that in the case of a being who performs a non-communicating cognitive act, e.g., the having of the silent thought that it is raining, the object of his act is structurally analogous to the sentence 'It is raining.'[10] In particular, in silent cognitive acts that include references to particulars, these references can be put in one-to-one correspondence with the types and kinds of singular-referring expressions. Thus,

---

[7] For a penetrating defense of the unanalyzability and theoretical indispensability of indexical reference, see P. F. Strawson, *Individuals* (London, Methuen & Co., 1959), pp. 117–120. See also the very interesting defense of indicators as singular terms by Manley Thompson in "On the Elimination of Singular Terms," *Mind*, vol. 67 (1959), pp. 361–376, against Quine's thesis that all singular terms are eliminable. I hasten to point out that my theses (I.6)–(I.7) do not seem incompatible with Quine's. I am not (here at any rate) laying claim to the necessity of the unanalyzable indexical roles being performed by singular terms in Quine's sense, and Quine, for his part, does not claim that all indicators can be defined or analyzed away. Quine lets 'now,' 'then,' 'here,' and 'there' to remain uneliminated as general terms (*op. cit.*, p. 185), and he also allows that the attributive use of these indicators as general terms depends on pointing. If Quine's reduction holds, my claims in (I.6) and (I.7) must be construed as claims about the indexical general terms that Quine introduces.

(I.9) To attribute a cognitive or linguistic propositional act to a person may very well be to attribute to the person in question purely indexical references.

Some cognitive relations are dispositional. A sleeping man, for example, may believe or know that the Earth is not flat. Here the disposition consists of a propensity to perform acts whose objects are analogous to sentences. We shall say that attributing a cognitive disposition to a person is to attribute to him *implicit* uses of sentences or words of certain types (without this implying a reduction of cognitive acts to linguistic acts). Now, a cognitive disposition may be exercised through different cognitive acts at very different times and places. Hence, to attribute a cognitive disposition to a person *P* is to attribute to *P* implicit uses of an indicator *D*, if and only if every possible exercise of *P*'s disposition is unavoidably a use of *D* by *P*. By (I.5) and (I.6), this is wholly the case for *P*'s use of the first-person pronoun. In brief,

(I.10) To attribute a cognitive disposition to a person *P* is to attribute to *P* implicit uses of an indicator *D*, only if *D* refers to a relatum, or place related to the place which is a relatum, of the cognitive acts through which the disposition is exercisable.

4. *Indicators in oratio obliqua.* By (I.1)–(I.2) a use of a genuine indicator refers directly to an entity, and does not depend for its reference on another expression. For example, the word 'he' is genuinely an indicator when it is used demonstratively in place of 'this' or 'that,' and it is an indexical description when it is short for 'this (that) man,' or something like 'that man with a beard.' 'He' is not indexical at all in the sentence 'Anthony came, after he called'; for here the word 'he' refers back to 'Anthony' and does not directly and by itself refer to Anthony. Similarly, the word 'I' is not an indicator in 'Mary told me that I am next in line.' With this sentence the speaker refers directly to himself by means of 'me'; the word 'I' in this case refers back to 'me.' Let us say that an *antecedent* of a pronoun or adverb in a sentence or clause is an expression of the sentence or clause to which the pronoun or adverb refers back. Hence,

(I.11) Whether in *oratio recta* or in *oratio obliqua*, (genuine) indicators have no antecedents.

Suppose now that Privatus asserts of a dead friend of his:

(2) Once it occurred to Jones that I buried a letter here.

Privatus' statement is philosophically perplexing. It seems to attribute to Jones demonstrative references by means of 'I,' and 'here'; yet, quite palpably, neither could Jones ever refer to Privatus in the first-person way nor was he at the time of Privatus' assertions in a position to make any demonstrative reference at all. By (I.1), the demonstrative references of (2) are Privatus', not Jones's. Thus, we have a philosophical question: What exactly is the proposition that, according to Privatu, Jones once took to be true? That is, what exactly are the types of the references that, according to Privatus, Jones made to Privatus and to the place in question?

One thing is clear. In spite of their misleading position in the *oratio obliqua* of (2), the indicators of (2) serve to mark the positions occupied by some unspecified referring expressions in the sentence formulating the unspecified proposition that, according to (2), Jones once took to be true. Those referring expressions that Jones used were, of course, either

---

[8] I myself have argued for this possibility in "*Lenguaje, Pensamiento y Realidad*," *Humanitas* (University of Nuevo León, Monterrey, Mexico), vol. 3 (1962), pp. 199–217; I touch on the issue in "The Private Language Argument," entry of *The Encyclopedia of Philosophy* (New York, Macmillan, forthcoming).

[9] This parallelism does not require that there be a component of a proposition *P* for every word or phrase of a sentence formulating *P*. The parallelism we need to accept is compatible with certain words or phrases being, e.g., mere grouping devices or scope signals like 'it is the case that.'

[10] For theories about thinking based on the structural analogy between a proposition and the sentences expressing it, see Wilfrid Sellars, "Empiricism and the Philosophy of Mind," in *Minnesota Studies in the Philosophy of Science*, vol. 1 (Minneapolis, University of Minnesota Press, 1956), pp. 305–328, and P. T. Geach, *Mental Acts* (London, Routledge & Kegan Paul, 1956), chs. 14, 17, and 22.

(a) single indicators; or
(b) names; or
(c) indexical descriptions, like 'this man,' 'my friend,' or 'five years ago today (now)'; or
(d) Leibnizian descriptions, i.e., descriptions that contain no indicators.

The actual proposition that, by (2), Jones once took to be true can be one of eight different types, depending on which sort of reference Jones made to Privatus and the place in question. And now we must raise another question: Is Privatus' statement (2) definite enough on this point? And the answer seems to be that it is not: Privatus' statement (2) is simply the statement to the effect that *one* of the eight types of propositions allowed by the two positions occupied by indicators was taken by Jones to be true. That is, Privatus' statement (2) is to be conceived of as a disjunction of certain statements which we proceed to identify.[11]

So far, we know that the indicators of (2) are in a sense misplaced, and that in their positions Jones's references must be represented. This suggests that the preliminary scheme of the analysis of (2) is this:

(2′) Once it occurred to Jones about me and about this place that $\underline{\qquad}_{\text{me}}$ buried a letter in $\underline{\qquad\qquad}_{\text{this place}}$,

where the subscripts indicate the antecedents of the expressions in the blanks.

Suppose that Privatus wanted to attribute to Jones purely indexical references. Hence, what we should put in the blanks of (2′) are precisely the quasi-indicators corresponding to the first-person and the place indicators. Thus, Privatus' statement would be more perspicuously put as

(3) Once it occurred to Jones about me and about this place that I* buried a letter here*.

Since we are mainly concerned with the typical logic of indexical reference, we may here lump together names and descriptions. Now, if Privatus wanted to attribute to Jones references by means of either names or descriptions, he would have had a hard time finding an appropriate idiom or locution in ordinary language. We shall, however, introduce a counterpart to quasi-indicators by means of a subscripted '*d*'. Thus Privatus' statement analogous to (2), except that through it Privatus attributes to Jones references by means of names or descriptions, can be formulated as

(4) Once it occurred to Jones about me and about this place that I_d had buried a letter here$_d$.

Thus, the first step in the analysis of (2) is to take it as the disjunction of (3), (4), and the two other statements "Once it occurred . . . that I* here$_d$" and "Once it occurred . . . that I_d . . . here*."

Since names and descriptions, Leibnizian or indexical, attribute properties to the entities they refer to, it is natural to think that (4) is analyzable as

(4N) There are properties ϕ-ness and μ-ness such that: I am the only ϕ, this place (here) is the only μ, and once it occurred to Jones that the only ϕ buried a letter in the only μ.

There is a general way in which (4N) is the analysis of (4), namely, the general way in which when a man says "I am ill" other people transmit his information by saying "He is ill" or even "The man . . . is ill." This is the way in which there is one and the same property that both Smith and Jones attribute to Privatus when Jones identifies Privatus by the description 'the man I met three days ago' and Smith identifies Privatus by the description 'the man Jones met three days ago.' On the one hand these descriptions identify Privatus by means of the same property, considered in itself, independently of the ways it is conceived of by Jones or Smith or anybody else. On the other hand, each of these descriptions ascribes to Privatus that property, not merely as it is in itself, but as it appears to either Jones or Smith: 'the man I met three days ago' ascribes to Privatus that property from Privatus' first-person perspective, whereas 'the man Jones met three days ago' ascribes to Privatus the same property from a third-person perspective. By (I.5)–(I.7) we know that a first-person way of considering a certain property is irreducible to a third-person way of considering it, and clearly the latter is irreducible to the former. Since each of the descriptions explicitly mentions the perspective from which the given property is attributed, each description really attributes a property-*cum*-perspective. Thus, in the general sense that a predicate denotes an attributed property, we can, and must distinguish between a

---

[11] This analysis of (2) occurred to me after George Nakhnikian pointed out an error in an earlier analysans candidate.

neutral, or *non-perspectival* property and the *perspectival* properties through which the former is presented. (I owe this terminology to W. Sellars.) I will use the italicized word '*properties*' to refer to perspectival properties, and I will indicate with subscripts the person to whose perspective of the world a given *property* belongs. The unitalicized word 'property' will be used to refer to neutral properties, or to neutral and perspectival properties indeterminately.

Since a person has to consider (neutral) properties in some way or other, we are never really dealing with properties naked, so to speak, of their perspectives. But as long as we deal with so-called extensional contexts only, we can enjoy the illusion that we are intellectually manipulating bare neutral properties. However, as soon as we move to the level of discourse about cognitive and linguistic relations, we must face up to the fact that we are in a partial egocentric predicament of properties.

This predicament creates the fundamental problem of the logic of communication: How to correlate another person's *properties* with our own and pass information across these correlations. The fundamental assumption regarding the possibility of communication is simply that such correlations can be achieved. And the exercise we are engaged in, viz., to determine the type of proposition that, according to Privatus, Jones once took to be true, is precisely an exercise in formulating some of the principles that allow the achievement of the communicating correlations.

In short, (4N) is, correctly, the *non-perspectival analysis* of (4). But since (4) is a non-extensional context, (4N) does not allow us to get hold of the precise statement that, according to Privatus, Jones once took to be true. The characteristic feature of (4N) is that one may know what properties ϕ_1-ness and μ_1-ness one should instantiate the quantifiers of (4N) into, without being in a position to produce the instance of (4N) that would be the straight report of it having occurred to Jones that such and such. To illustrate, suppose that

ϕ_1-ness = man-met-by-Jones-on-July-1-1960-ness

μ_1-ness = flowerbed-dug-by-Jones-ness

That is, let Privatus know that:

(5) There are properties ϕ-ness and μ-ness such: ϕ-ness = man-met-by-Jones-on-July-1-1960-ness, μ-ness = flowerbed-dug-by-Jones-ness, I am the only person who is ϕ, this place is the only one which is μ, and once it occurred to Jones that the only person who is ϕ buried a letter in the only place which is μ.

From (5) Privatus cannot conclude that:

(6) I am the only man met by Jones on July 1, 1960, this place is the only flowerbed dug by Jones, and once it occurred to Jones that the only man met by Jones on July 1, 1960 buried a letter in the only flowerbed dug by Jones.

The difficulty lies in the last conjunct "once it occurred to Jones that the only man met by Jones on July 1, 1960 buried a letter in the only flowerbed dug by Jones." Obviously, this may be false even though Jones did refer to Privatus and the place in question as possessors of these properties ϕ_1-ness and μ_1-ness when it occurred to him what it occurred to him according to Privatus' statement (2). For instance, it may have occurred to Jones what he would have reported by saying "The only man met by me on July 1, 1960 buried a letter in the only flowerbed dug by me." The crux of the matter is that in a proposition that a person takes to be true, or simply entertains, a property enters not merely as a neutral property, but as a property determined by the way in which the person in question grasps it or conceives of it. Thus, while Privatus' statement (5) presents the above properties ϕ_1-ness and μ_1-ness as they are grasped by Privatus, his statement (6) presents ϕ_1-ness and μ_1-ness both as grasped by Privatus, in the oo-prefix, and as grasped by Jones, in the *oratio obliqua*. This is why the inference from (5) to (6) is invalid.

Let us turn, then, to *properties*. Consider an expression containing no indicators or quasi-indicators that formulates a property. Clearly, there are four ways of considering or predicating the property in question:

(i) the first-person way;
(ii) the second-person way;
(iii) the third-person way involving third-person indicators;
(iv) the third-person way formulated by the given expression.

Let us use subscripts to signal these ways of considering properties, as follows: '$_1$' for (i), '$_2$' for (ii), '$_i$' for (iii), and '$_o$' for (iv). Thus, the property being-met-by-Jones-5-days-before-July-1-1965 determines the following *properties* from Jones's perspective, to be called *properties*$_J$:

(a) being a person met by Jones$_1$ 5 days before (July 1, 1965)$_i$;
(b) being met by Jones$_1$ 5 days before (July 1, 1965)$_o$;

(c) being met by Jones$_o$ 5 days before (July 1, 1965)$_i$; and
(d) being met by Jones$_o$ 5 days before (July 1, 1965)$_o$.

On the other hand, for Privatus it determines the *properties*$_P$:

(d) as above;
(e) being a person met by Jones$_2$ 5 days before (July 1, 1965)$_i$;
(f) being a person met by Jones$_1$ 5 days before (July 1, 1965)$_o$;
(g) being a person met by Jones$_o$ 5 days before (July 1, 1965)$_i$; etc.

Note that, by (I.1) and (I.8), *properties* (c) and (g) are quite different.

By our definition of quasi-indicators we know that quasi-indicators represent the uses of indicators in *oratio obliqua*. Thus, when a *property* ϕ_X-ness appears in *oratio obliqua*, we can represent it by means of an expression which results by putting quasi-indicators instead of the indicators in the expression denoting ϕ_X-ness. The resulting expression does not, of course, represent ϕ_X-ness *simpliciter*, but as it is considered by the one who makes the statement, and even perhaps as it is considered by a few other persons as well.

In general, let $P(i_1, \ldots, i_n)$ be a phrase that Jones uses (or can use) on a certain occasion O to formulate a *property*$_J$, where $i_1, \ldots, i_n$ are indicators that Jones uses (or can use) on O to refer to entities $E_1, E_2, \ldots, E_n$. Let $P(i_1/e_1, \ldots, i_n/e_n)$ be the phrase that results from $P(i_1, \ldots, i_n)$ by replacing each indicator i_k by the corresponding expression e_k, which is an expression that a man X uses (or can use) on O to refer to entity E_k. Clearly, $E(i_1/e_1, \ldots, i_n/e_n)$ formulates a *property*$_X$, and to indicate its correspondence with the *property*$_J$ formulated by $P(i_1, \ldots i_n)$, we shall call it a *property*$_{J/X}$. Similarly, let $P(i_1/q_1, \ldots, i_n/q_n)$ be the expression that results from $P(i_1, \ldots, i_n)$ by replacing each i_k by its corresponding third-person quasi-indicator q_k. In the next section we shall discuss this correspondence between indicators and quasi-indicators in detail. That there is such a correspondence is obvious from the fact that Privatus' statements (2) and (4) are not addressed to Jones, but are about Jones having taken a certain statement to be true, so that Privatus must have some technique (namely, the use of quasi-indicators) for representing in a third-person way Jones's uses of indicators (if any). We shall say that $P(i_1/q_1, \ldots, i_n/q_n)$ formulates a *property*$_{*J}$, i.e., formulates someone's, here Privatus', *oratio-obliqua* way of considering the *property*$_J$ in question. Since the quasi-indicators q_k have nothing to do with Privatus, the *property*$_{*J}$ is really an inter-personal or inter-subjective version of the corresponding *property*$_J$, and while by (I.7) most *properties*$_J$ must yield even for Jones to other *properties*$_J$, the corresponding *properties*$_{*J}$ are permanent: anybody can refer to them at any time, as we shall see in detail later on.

With this machinery we can give a deeper analysis of Privatus' statement (4):

(4) Once it occurred to Jones about me and about this place that I$_d$ buried a letter here$_d$.

A perspectival analysis in terms of *properties* is as follows:

(4P) There are *properties* ϕ_J-ness and μ_J-ness such that: I am the only person who is $\phi_{J/I}$, this place is the only place which is $\mu_{J/I}$, and once it occurred to Jones that the only person who is ϕ_{*J} buried a letter in the only place which is μ_{*J}.

Note that since Privatus is in (4) referring to himself in the first-person way, the *property*$_P$ related to ϕ_J-ness and μ_J-ness must yield the first-person schematic symbols '$\phi_{J/I}$' and '$\mu_{J/I}$' representing properties and having adjectives as substituends.

Consider now Privatus' statement

(7) Jones believed that I buried a letter here.

By (7) Privatus attributes to Jones a cognitive disposition and, hence, only implicit references to him and to the place in question. These references are not indexical, even if Privatus' statement (7) reports what Jones believed at a single instant. Indeed, suppose that at a certain instant t it occurred to Jones out loud: "He [pointing to Privatus] buried a letter here [pointing to the place in question]," and suppose that Jones died at that very instant. In this case, Jones never exercised the disposition which is his belief in anything but thoughts containing indexical references. Yet, since these references are by (I.7) eliminable for Jones, the onsetting of his belief is precisely the onsetting of procedures of elimination, which were never exercised, but were nevertheless acquired. Furthermore, even at instant t Jones had, dispositionally, i.e., as part of his belief, ways of referring to the place and to Privatus other than by the use of single indicators. For instance, he could have referred to Privatus by means of an indexical description such as 'this man,' 'the man over there,' 'the person I see on my right,' etc. Thus, I submit that in the case of (7) the non-perspectival and the perspectival analyses are non-disjunctive, as follows:

(7N) There are properties ϕ-ness and μ-ness such that: I am the only person who is ϕ, this place is the only place which is μ, and Jones believed that the only person who is ϕ buried a letter in the only place which is μ;

(7P) There are *properties* ϕ_J-ness, μ_J-ness, $\phi_{J/I}$-ness, and $\mu_{J/I}$-ness such that: I am the only person who is $\phi_{J/I}$, this place is the only place which is $\mu_{J/I}$, and Jones believed that the only person who is ϕ_{*J} buried a letter in the only place which is μ_{*J}.

Evidently, the discussion of Privatus' statements (2)–(7) is applicable, *mutatis mutandis*, not only to other statements by Privatus, but to any statement made by any other person. Thus, we shall regard as established that:

(I.12) Indicators in *oratio obliqua* are eliminable for their users in terms of indicators in *oratio recta;*

(I.13) Indicators in *oratio obliqua* representing the propositional relata of cognitive or linguistic acts are analyzable non-perspectivally as well as perspectivally. The procedures of analysis can be carried out one step at a time for each oo-prefix, as illustrated by (4N) and (4P), respectively;

(I.14) Indicators in *oratio obliqua* representing propositional relata of cognitive dispositions are analyzable non-perspectivally and perspectivally. The procedures of analysis can be carried out one step at a time for each oo-prefix, as is illustrated by the analysis of (7) as (7N) and as (7P), respectively.

5. *Quasi-indicators.* The preceding discussion of indicators is not only a good background for, but has to be complemented by, the study of quasi-indicators.

(Q.1) Quasi-indicators do not make demonstrative references. They may even fail to make reference to single entities, for *they can play the role of variables of quantification.* For instance, in "Always everywhere a boy tells a girl that he* will love her* from then* on" the quasi-indicators 'he*,' 'her*,' and 'then*' are (also) variables bound, respectively, by the quantifiers 'a boy,' 'a girl,' and 'always.'

(Q.2) Quasi-indicators have necessarily an antecedent to which they refer back, but they are not replaceable by their antecedents. Quasi-indicators are both referentially and syntactically dependent: they are syncategorematic expressions. And the sentences (clauses) that contain quasi-indicators without containing the latter's antecedents are also syncategorematic. An interesting example is provided by the following inference:

(8) While in Brown's office Jones knew that Smith was killed there*

hence

(9) Smith was killed there.

It is obviously valid, and sentence (9) expresses a complete statement in *oratio recta*, in which its 'there' can be replaced by its antecedent 'in Brown's office.' Hence, by (Q.2) 'there' in (9) is not an instance of the quasi-indicator 'there*.'

It might be thought that 'there' in (9) is precisely another occurrence of the word 'there*' of the premiss. It has to be the same quasi-indicator, it might be argued, since the above inference is validated by the rule.

(K.1) From a statement of the form "*X* knows that *p*" you may infer that *p*.

This argument is, however, invalid: (K.1) says nothing about detaching the sentence to the right of 'knows that.' (I have shown in *He*[12] that (K.1) itself is a rule that cannot always be applied.) What the argument needs is a rule like

(K.2) If a sentence of the form '*X* knows that *p*' formulates a statement that you accept, then you may detach the sentence (or clause) represented by '*p*' and use it by itself, with no change of meaning or sense, to make the very same statement which it formulates as a part of the longer sentence '*X* knows that *p*.'

The word 'there' in (9) is, of course, an instance of the quasi-indicator 'there*,' which does appear in (8), if the inference "(8), hence (9)" is validated by (K.2).

Suppose, then, that Privatus believed (8) to be true, that while he himself was in Brown's office he made out loud that inference, and that

(10) While in Brown's office Privatus did not believe that Brown's office was there*.

While drawing the inference from (8) to (9) Privatus believed that both (8) and (9), i.e., he believed that (while in Brown's office Jones knew that Smith was killed there* and Smith was killed there). Make now the assumption that the word 'there' in the conclusion (9) is the quasi-indicator 'there*.' Then since "*X* believes that *p* and *q*" entails "*X* believes that *q*,"

(10′) While in Brown's office Privatus believed that Brown's office was there*.

Since (10′) contradicts (10), the assumption is false. Since the use of 'there' in (9) as part of the inference "(8), hence (9)" is the same regardless of the truth of (10), we conclude that in no case is 'there' in (9) the quasi-indicator 'there*.' Therefore, the inference "(8), hence (9)" is not validated by rule (K.2), and (K.2) is *not* a universally valid rule.[13]

---

[12] Toward the end of Section 4.

(Q.3) Quasi-indicators have an epistemological priority over indicators and indexical descriptions: a person's uses of indicators are not only eliminable for another person in terms of quasi-indicators, but the latter are the only linguistic expressions that preserve the full force of the former. This is clear from the discussion in section 4 above.

(Q.4) There are two types of quasi-indicators: (A) those, to be called *primary quasi-indicators*, whose antecedents belong in a strict oo-prefix to which they are subordinated, and (B) those, to be called *secondary quasi-indicators*, whose antecedents lie in no strict, but only in an extended oo-prefix. The former's antecedents denote structural elements of cognitive or linguistic acts, while the latter's antecedents denote entities external to cognitive or linguistic relations. The distinction is epistemologically important, as we shall see in (Q.8)–(Q.10). For example, suppose that Jones grabbed at time *t* the sharpest knife in the kitchen and said to Mary:

(11) I am going to kill you with this here now!

Privatus can naturally report the proceedings by asserting:

(12) Jones grabbed at *t* the sharpest knife in the kitchen and said to Mary that he was going to kill her with it there then.

From (11) and (12) we see that the word 'it' in (12) represents an actual use of the indicator 'this' by Jones, and it has as antecedent the occurrence of the phrase 'the sharpest knife in the kitchen' in (12); it is the quasi-indicator 'it*.' But the antecedent of 'it' is not a part of, while the antecedents of the quasi-indicators 'he,' 'her,' 'there,' and 'then' are parts of, the strict oo-prefix of (12).

Sentence (12) by itself contains a subtle ambiguity, analogous to the one we found in sentence (2) above. It can be used to make three different kinds of statement, depending on whether its 'it' is used to attribute to Jones a reference to the sharpest knife in question, which is (i) purely indexical, or (ii) descriptive, or (iii) one or the other, indeterminately. With the notation already introduced we can formulate these distinctions perspicuously as follows, where the dots indicate that the remaining parts of the sentence (12) are preserved:

(12.i) . . . it_{*J} . . .;
(12.ii) . . . it_d . . .;
(12.iii) Either (12.i) or (12.ii).

With the background furnished by statement (11) we know that statement (12) is the same as statement (12.i). As we shall see later on, (Q.8)–(Q.10), the quasi-indicator 'it*,' hence 'it_{*J},' is unanalyzable. On the other hand, (12.ii) and (12.iii) are analyzable along the lines of the procedures illustrated in section 4 for the analysis of indicators in *oratio obliqua*.

By definition, 'it_d' in (12) expresses that by means of some description, indexical or otherwise, Jones referred to the knife referred to by Privatus as the sharpest one in a certain kitchen. Jones's description may, of course, be entirely different from Privatus'. In brief, we have here very much the same situation as we had in section 4 for the case of indicators. Thus, we also have a non-perspectival and a perspectival analysis of Privatus' statement (12.ii) as follows:

(12.iiN) Jones grabbed the sharpest knife in the kitchen, and there is a property ϕ-ness such that: that knife is the only thing which is ϕ, and Jones said to Mary that he* was going to kill her* with the only thing which is ϕ there* then*;

(12.iiP) Jones grabbed the sharpest knife in the kitchen, and there is a *property* ϕ_J-ness such that: that knife is the only thing which is $\phi_{J/I}$, and Jones said to Mary that he* was going to kill her* with the only thing which is ϕ_{*J} there* then*.

I have put the asterisks signaling the quasi-indexical role of the pronouns they attach to, in conformity with the definition of quasi-indicators and principles (I.9) and (I.10).

(Q.5) The chart below both establishes the third-person correlation between indicators and quasi-indicators appealed to in (2P), (4P), (7P), and (12.iiP), and summarizes previous results, like (I.7), (I.9), (I.10). The columns contain the following:

1. the type of entity X a person P refers to;
2. the indicator I by means of which P refers to X;
3. the type of second-order relation R that has

---

[13] Obviously a similar counter-example against (K.2) can be produced using 'then*' and 'it*.' In *He*, Section 4, appears the original argument of this type employing 'he*' qua representation of uses of the indicator 'I'.

as its relata P and the statement constituted by P's mentioned reference to X by means of I;

4. the eventual type of relation, i.e., disposition or occurrence, of R;
5. the third-person quasi-indicator Q that corresponds to I; and
6. the type of oo-prefix containing the antecedent of Q.

We use the following abbreviations:

C–cognitive relation; *L*–linguistic relation;
O–occurrence; *D*–disposition;
S–strict oo-prefix; *E*–extended oo-prefix

1	2	3	4	5	6
I. Self	first-person pronoun: 'I'	C, L	O, D	'he*'	S
II. Time	adverb of time 'now' (mainly)	C, L	$O, \frac{1}{2}D$	'then*'	S
III. Space:					
(a) near	adverb of place 'here' (mainly)	C, L	$O, \frac{1}{2}D^1$	'there*' 'here*'	S
(b) far	adverb of place 'there' (mainly)	C, L	O	'there*'	E
IV. Other Selves:					
(a) addressed to	second-person pronoun: 'you'	L	O	'he*'	S
(b) others	third-person pronoun: 'he'	C, L	O	'he*'	E
V. Objects	demonstrative pronouns: 'this,' 'that'	C, L	O	'it*'	E

In this table, we use $\frac{1}{2}D$ (half disposition) because implicit references to indicators are limited as required by (I.10). Moreover, it should be noted that the quasi-indicator 'now*' has as antecedent the indicator 'now,' and similarly the quasi-indicator 'here*' has as antecedent the indicator 'here.'

By putting 'I*' instead of 'he*' in the next to the last column we can produce the first-person correlation between indicators and quasi-indicators, where 'I*' has the indicator 'I' as antecedent. Similarly, by putting 'you*' in place of 'he*' we obtain the second-person correlation, where, again, 'you*' has the indicator 'you' as antecedent.

The above table makes it apparent how the five (or seven) types of indicators compare with one another. It shows that each type of indicator is fully characterized by pairing a category mentioned in the first column with a pattern yielded by third, fourth, and last columns. For example, type V of indicator is fully determined by the pair (object: C–L–O–E); the first-person indicator[14] is fully determined by the pair (self: C–L–O–D–S). In these pairs, the category determines the only amount of classifying that indicators perform, and this by implication only, indirectly, so to speak, not by attribution. The pattern, on the other hand, merely indicates how wide is the range of use of an indicator. From this we may see immediately that the first-person indicator cannot be defined in terms of the other indicators: no combination of the latter has the wide range of the former.

At any rate, it is clear that

(I.15) The pairs of categories and patterns are both adequate conditions for defining the several types of indicator, and criteria for ascertaining the type of indicator to which a certain word of a given language belongs.

Let $S(E)$ be a sentence containing an occurrence O of an expression E, and let $S(E/E')$ be the sentence that results from $S(E)$ just by replacing O with an occurrence of E'. I shall say that O is *a proxy for* E' *in* $S(E)$, if and only if one and the same person using $S(E)$ assertively on a given occasion would make the very same statement that he would formulate by using $S(E/E')$ assertively on that occasion. I shall say that O is *a substitute for* E' *in* $S(E)$, if and only if the statement any person would formulate on a certain occasion by using $S(E)$ assertively is logically equivalent to the statement the same person would formulate by using $S(E/E')$ assertively on the very same occasion.

Quasi-indicators are substitutes for other referring expressions in some trivial cases, namely, in sentences formulating self-contradictions, necessary truths, or certain redundant statements. For instance, "—— & Peter believes that he* is ϕ, or ——" is tautologically equivalent to "——"; and replacing the explicit occurrence of 'he*' with any other referring expression does not destroy the equivalence.

(Q.6) An occurrence of a quasi-indicator is not a proxy or a substitute for an indicator or indexical description in some sentence of the form '$L[x, y_1, \ldots, y_n, p, t]$ (——)' or '$\sim L[x, y_1, \ldots, y_n, p, t]$ (——)'. This is obvious from our discussion of indicators in *oratio obliqua*.[15]

(Q.7) An occurrence of a quasi-indicator is not a proxy or a substitute for a name or a description containing no quasi-indicators, in some sentence of the form '$L[x, y_1, \ldots, y_n, p, t]$ (—)', or the form '$\sim L[x, y, \ldots, y_n, p, t]$ (—)'.

---

[14] Note that the first-person pronoun, which is also used as a quasi-indicator, is not being characterized here.

[15] In *He*, Section 2, I examine several natural ways in which it might be thought that the primary indicator 'he*' could be construed as a vehicle for indexical reference.

We can assimilate reference by means of names to reference by means of descriptions. On the one hand, many names apply to several objects, so that to guarantee uniqueness of reference they must be supplemented with a description. On the other hand, one acquires (or learns) a name for an entity either through a description or through a personal confrontation that gives a descriptional support to the name. (This assimilation is not necessary, but it simplifies the argument. The important thing is that reference by means of a name is not equivalent to reference by an indicator.) By (Q.6) we need not consider the case of descriptions containing indicators.

The following statement verifies (Q.7):

(13) On May 15, 1911, the German Emperor believed that it was raining then*.

Clearly, the German Emperor may have had no idea at all as to the date on which it was raining then*, and he may have also failed to have any other non-indexical description that uniquely characterized May 15, 1911. I cannot, however, muster a formal argument to show this. But a general argument for (Q.7) can be framed with the help of (I.3).

Let I be an indicator. Then by (I.3), there is at least one sentence $S(I)$ such that for no description D containing neither indicators nor quasi-indicators, the statement "$S(I)$" formulated by $S(I)$, if $S(I)$ is used assertively in circumstances C by some person X to refer to an object O, is not equivalent to the statement "$S(I/D)$" that $S(I/D)$ would formulate if assertively used by X in C to refer to O. Thus the statements "$S(I/D)$" and "$S(I)$" are both self-consistent. Now, let Q be the quasi-indicator that represents uses of I in *oratio obliqua*. Then by (Q.2), the clause $S(I/Q)$ is syncategorematic and expresses the statement "$S(I)$" only in a context of the form '$L[x, y_1, \ldots, y_n, p, t]$ $(S(I/Q))$' or even a more inclusive one. Consider one of the statements formulated by a sentence of this form and call it "$OS(I/Q)$." This is a statement to the effect that the person X stands at some place p and time t in some cognitive (or linguistic) relation R to (some addressees and) the statement "$S(I)$." On the other hand, the corresponding sentence of the form '$L[x, y_1, \ldots, y_n, p, t]$ $(S(I/D))$' formulates the statement, "$OS(I/D)$," that X stands at the same place p and time t in the same relation R to (the same addressees and) the statement "$S(I/D)$." Obviously, we can choose a name or some self-consistent description of the person X to put in the place of 'x' in the oo-prefix of the above sentences. Thus, we can choose statements "$OS(I/D)$" and "$OS(I/Q)$," which are contingent. Hence, by the principle (P) below it follows that the statement "$OS(I/Q)$" is not equivalent to the statement "$OS(I/D)$," for whatever description D that contains neither quasi-indicators nor indicators:

(P) If a statement formulated by a sentence of the form '$L[x, y_1, \ldots, y_n, p, t]$ (—)' is contingent, then if it is equivalent to the corresponding statement of the form '$L[x, y_1, \ldots, y_n, p, t]$ (. . .)', then the statement formulated by what fills the blank '—' is equivalent to the statement formulated by what fills the blank '—' provided both are self-consistent.

By (P), to believe that a certain proposition S is true is equivalent to believing that a certain proposition S' is true, only if S and S' are themselves logically equivalent. For example, believing that it is both raining and hailing is equivalent to believing that it is both hailing and raining, and, clearly, "it is raining and hailing" is equivalent to "it is hailing and raining."

It should be noted that if in the above reasoning the sentence expressing "$OS(I/Q)$" is categorematic, then Q is a primary quasi-indicator. On the other hand, if the sentence in question is syncategorematic, i.e., it formulates "$OS(I/Q)$" only by being concatenated with another clause, with which it forms a larger context C, then Q may be a primary or a secondary quasi-indicator. In the latter case (as we shall see later on) Q may be a substitute for some description variable in C. But this in no way affects its being substitute for no description containing no quasi-indicators in the smaller context that formulates "$OS(I/Q)$."

If "$OS(I/Q)$" is not equivalent to "$OS(I/D)$," then "$\sim OS(I/Q)$" is not equivalent to "$\sim OS(I/D)$." And since being a proxy entails being a substitute for the same expression in the same sentence, (Q.7) is fully established by the preceding argument.

6. *Unanalyzability of strict quasi-indicators.* Variables of quantification are unanalyzable referring expressions. Thus, inasmuch as quasi-indicators sometimes function as variables of quantification, they cannot be analyzed in terms of names, indicators, or descriptions containing no quasi-indicators. But from (Q.6) and (Q.7) the stronger result follows that no quasi-indicator referring to a single entity is analyzable solely in terms of

indicators, names, or descriptions (indexical or Leibnizian) that do not contain quasi-indicators. Furthermore the procedures discussed in section 4 for the analysis of indicators in *oratio obliqua* cannot produce an analysis of quasi-indicators. Consider, for example, Privatus' statement

(14) Jones thought at 3 p.m. yesterday that Paul was then* sick.

If we apply to (14) the procedure for non-perspectival analysis we find

(14N) There is a property ϕ-ness such that: the (time) $\phi = 3$ p.m. yesterday and Jones thought at 3 p.m. yesterday that Paul was sick at the (time) ϕ.

But (14N) and (14) can have different truth-value. For while (14) attributes to Jones a purely indexical reference to the time of his thinking, (14N) attributes to him a descriptive reference.

If we apply the procedure for a perspectival analysis we obtain:

(14P) There is a *property* ϕ_J-ness such that: the time which is the only $\phi_{J_1} = 3$ p.m. yesterday, and Jones thought at 3 p.m. yesterday that Paul was sick at the only time which is ϕ_{*J}.

Again, (14) and (14P) can have different truth-values: it is possible that when Jones thought what Privatus reports, he just made either a purely indexical or a descriptive reference to the time of his thinking, but not both. In the case of (14P) there is the additional point that even if it were an analysis of the singular quasi-indicator 'then*' of (14), it would be an analysis in terms of the quasi-indexical variables 'ϕ_{*J}.' Thus, the perspectival procedures would constitute at best an analysis of singular quasi-indicators.

Now, a language or conceptual scheme through which the only self or person that can be conceived is the one who uses the scheme, obviously, does not allow for the conception of the contrast between two persons. Such a language does not allow for the contrast between first- and other-person statements, or between second- and other-person statements. Thus, whatever verbal and whatever pronominal forms are part of that language, they will simply fail to express the distinction among grammatical persons. Thus, we may suppose, for concreteness, that in such a language we only have what appears to be first-person statements of the form "I am ϕ." Yet the pronoun 'I' cannot refer in such statements to an entity which has general mental properties, i.e., properties that do not necessarily have just one instance. For a language to have predicates that formulate such general properties is for it to allow the framing of statements which formulate the possibility that these properties be exemplified by several entities. Thus, the pronoun 'I' in the seemingly first-person statements of the type of language we are discussing has to refer to an entity of a category, namely, self, for which the language does not allow other members. Hence, I submit that the seeming first-person statements in question are really either (i) disguised strictly impersonal statements (like "There is a pain here," "There was a thought that it will rain," and "There was a belief that it would rain now"), or (ii) statements in which the entity the word 'I' refers to is simply the totality (and hence the only entity of its sort) of all the experiences, thoughts, etc., that constitute the user of the conceptual scheme in question.[16] I shall assume that perceptual thinking necessarily involves demonstrative reference.

On the other hand, a conceptual scheme that allows for the mere possibility of other thinking beings, which have perceptions but need not be aware of themselves, allows necessarily for the contrast either between first- and other-person statements, or between third-person statements about one entity and third-person statements about another entity. That is, the user of the scheme can at least consider statements like "This is a cow and *X* thinks that it* is pregnant," which attribute to the other thinking beings experiences and experiential judgments that involve the use of demonstrative reference. In brief,

(Q.8) A conceptual scheme that allows for the possibility of other perceiving thinking beings (even though their reality isn't affirmed in the scheme) is characterized by the possession of the unique and unanalyzable referring mechanism constituted by the secondary quasi-indicators discussed in (Q.4).

A person is not only a center of consciousness or a factory of thoughts and reasonings. A person is a center of self-consciousness. Thus, a conceptual scheme that allows even for the mere possibility of several persons (or self-aware selves) necessarily allows for: (i) the formulation of the contrast between first- and third-person statements, and (ii) the formulation of statements of the form "*X* thinks at place *p* at time *t* that he* is —— there* then*" and of the form "I think that I* am ——." Therefore

---

[16] For more details on the impersonal core of first-person statements, see P. T. Geach, *op. cit.*, pp. 117–121, and my "Behavior and Consciousness," in *Studies in the Philosophy of Mind* (Detroit, Wayne State University Press, 1966), Sections 9 and 10. In *Individuals*, p. 99n, Strawson formulates the much stronger principle that "the idea of a predicate is correlative with that of a *range* of *distinguishable* individuals" (my italics in 'distinguishable'), and I have criticized this principle in *La Dialéctica de la Consciencia de Sí Mismo* (Guatemala, University of San Carlos Press, 1960), pp. 45–50.

(I.16) A conceptual scheme that allows for the mere possibility of several persons (or self-aware selves) is characterized by the possession of the unanalyzable referring mechanism constituted by the first-person indicator, and

(Q.9) by the possession of the unanalyzable referring mechanism constituted by the primary first- and third-person quasi-indicators.

Finally, a conceptual scheme that allows for the possibility of communication necessarily allows for: (i) the formulation of the contrast between first- and second-person statements, and (ii) the formulation of second-person statements of the form "You believe that you* are ——." Hence,

(I.17) A conceptual scheme that allows for the possibility of communication is characterized by the possession of the unanalyzable referring mechanism constituted by the second-person indicator, and

(Q.10) by the possession of the unanalyzable referring mechanism constituted by the primary second-person quasi-indicator.

7. *Quasi-indicators in n-fold oratio obliqua.* We proceed to discuss the quasi-indicators' counterpart of indicators in *oratio obliqua*, namely, the case of quasi-indicators subordinated to oo-prefixes which themselves occur in *oratio obliqua*. To start with, suppose that Privatus asserts

(15) In Brown's office Smith thought that in Brown's office Jones thought that a treasure was hidden there*.

Then we have a problem: to ascertain what exactly was Privatus' statement. His utterance has a noteworthy ambiguity. It may be brought out neatly by imagining Smith to be in Brown's office making the first-person statement corresponding to (15). He has a choice among:

(16) Here I thought that in Brown's office Jones thought that a treasure was hidden here;

(17) Here I thought that in Brown's office Jones thought that a treasure was hidden there*;

and

(18) Here I thought that here Jones thought that a treasure was hidden here.

In short, the ambiguity lies in that the quasi-indicator 'there*' in (15) may have as its antecedent either the second occurrence of 'in Brown's office' to the left of 'there,' which is in *oratio recta*, or the other occurrence, which is in *oratio obliqua*. In the former case Privatus' statement is the counterpart of (16), while in the latter it is the counterpart of (17). But there seems to be no natural way of expressing this distinction in English. It can be expressed, however, by means of some artificial improvement upon English, e.g., by attaching to 'there*' a subscript that indicates the distance between it and its antecedent. Either the antecedent, as in the case of (15), lies in a strict oo-prefix *P*, or it lies in an extended oo-prefix *P*; thus, we can measure the distance between an occurrence of a quasi-indicator and its antecedent by counting the strict oo-prefixes up to *P* to which that occurrence is subordinated. For instance, Privatus' counterpart of (16) is, then:

(19) In Brown's office Smith thought that in Brown's office Jones thought that a treasure was hidden there$*_2$.

Privatus' statement which is the counterpart of (17) is:

(20) In Brown's office Smith thought that in Brown's office Jones thought that a treasure was hidden there$*_1$.

By (Q.2)–(Q.9), the 'there$*_1$' in (20) is not analyzable. But by (I.12) the indicator 'here' at the end of (16) is eliminable for the speaker, and by (I.10)–(I.11) the elimination conforms to the procedures illustrated in Section 4. Since 'there$*_2$' in (19) represents the use of 'here' in (16), then 'there$*_2$' is eliminable for the speaker by the same procedures illustrated in Section 4. Thus, Privatus' statement (19) is more perspicuously formulated as the perspectival statement:

(19P) In Brown's office Smith thought that there is a *property* ϕ_J-ness such that: the only place which is $\phi_{J/P}$ = Brown's office, and in Brown's office Jones thought that a treasure was hidden at the only place which is ϕ_{*J_1}.

Here I have put 'ϕ_{*J_1}' instead of merely 'ϕ_{*J}' to indicate that the quasi-indicators in the description which occupies the place of '$\phi*_{J_1}$' have as antecedents items in the description that occupies the place of 'ϕ_J.' It is easy to see that at this stage our subscript notation allows of a simplification (hereafter adopted) without diminishing its power to dissolve ambiguity. An expression of the form '$\phi*_{J_n}$' is a quasi-indicator such that (i) its antecedent lies *n* units to the left of it, in an oo-prefix

P, and (ii) it represents *J*'s uses of indicators, where '*J*' stands for an expression both referring to person *J* and being the subject of the verb in the same oo-prefix *P*. Thus, we can drop the subscript '*J*' from 'ϕ_{*J_n},' since by looking at the *n*th prefix to its right we can find both its antecedent and the expression '*J*' stands for.

Now, it is a trivial exercise to construct parallel distinctions for each of the other quasi-indicators.[17] And if we regard the subscript notation as a characteristic feature of quasi-indicators, we can, then, conclude that

(Q.11) Quasi-indicators that are more than one unit removed from their antecedents are eliminable. That is, quasi-indicators of distance 1 are unanalyzable, but all others are analyzable in terms of those of distance 1.

Our subscript notation is, however, not yet fully adequate to eliminate the ambiguities of some sentences containing quasi-indicators. Consider, for example, Privatus' statement:

(21) In Brown's office Jones thought that Smith claimed there* that a treasure was hidden there*.

Clearly, the first 'there*' has 'in Brown's office' as its antecedent, and Privatus is not, by (21) ascribing to Jones the knowledge or even the idea that the place at which he was when he had the thought in question was precisely Brown's office. But the 'there*' at the end of (21) is ambiguous. No doubt, in some sense it also has 'in Brown's office' as its antecedent. But whose uses of the indicator 'here' does the second 'there*' represent? If it represents Jones', then the second 'there*' has 'in Brown's office' as its direct antecedent. If it represents Smith's, then its *immediate* antecedent is the first 'there*' and 'in Brown's office' is only its *ultimate* antecedent. Yet there is no normal way of expressing this distinction in ordinary language.

At this juncture I want to propose a notation that can be used as an auxiliary to English to make all the necessary distinctions that quasi-indicators in *n*-fold *oratio obliqua* allow. This notation will also serve the theoretical purpose of allowing the precise formulation of a fragment of the logic of quasi-indicators.

8. *Notation for quasi-indicators*. First, we unify the expression of all quasi-indicators by dropping the ordinary pronouns, adverbs, etc., used to express them, and simply indicate the underlying quasi-indexical role by means of ' * '. Second, we prefix ' * ' to an occurrence *O* of the ultimate antecedent of the quasi-indicator whose expression we are framing. Third, we suffix to such occurrence *O* a subscript that indicates the distance between the quasi-indicator and its immediate antecedent. (We could drop ' * ' altogether and express the pure quasi-indexical role by the mere suffixation of the subscript. But we preserve ' * ' and suggest that it be read as 'he,' 'then,' 'it,' etc., as the case may be.)

With this notation we can unambiguously write the two interpretations of (21) as

(22) In Brown's office Jones thought that Smith claimed *(in Brown's office)$_1$ that a treasure was hidden *(in Brown's office)$_2$;

and

(23) In Brown's office Jones thought that Smith claimed *(in Brown's office)$_1$ that a treasure was buried *(in Brown's office)$_1$.

This exegesis of (21) satisfies (Q.10). Since the second 'there*' of (22) is at distance 2 from its antecedent, by (Q.10) it is analyzable, and its subscript '2' is an immediate indication of its analyzability. Similarly, the subscripts of (23) clearly convey that none of its quasi-indicators is analyzable.

Parallel examples show that the temporal quasi-indicator 'then*' has parallel ambiguities that can be dissolved by our notation. But the quasi-indicator 'he*,' when it represents a use of the indicator 'I', not, e.g., of the indicator 'this (man),' does not allow the ambiguity parallel to that of (21). Consider, for example

(24) Jones believes that he* believes that he* is sick.

If Jones were to make the first-person counterpart statement, he would say:

(25) I believe that I believe that I am sick.

In short, Privatus' statement (24) attributes to Jones belief in the truth of a first-person statement about somebody's being sick, at the same time that it attributes to Jones belief in the identity of that somebody to himself. Thus, in our notation (24) comes out as

(24A) Jones believes that *Jones$_1$ believes that *Jones$_1$ is sick.

Again, (24A) palpably shows that by (Q.10) neither occurrence of the quasi-indicator is analyzable. Here we have a distinctive feature of the primary quasi-indicator 'he*' that sets it apart from all the other quasi-indicators.

The secondary quasi-indicator 'he*,' which represents uses of the personal indicator 'this (that),' allows of an ambiguity analogous to the one of (21). For example,

---

[17] For the case of the primary indicator 'he*' see *He*, examples (18)–(20).

(26) John told of Mary that Brown would think of her* that she* had been appointed.

The occurrence of ‘she*’ expresses a demonstrative reference either by Brown, in which case it is ‘$*(\text{Mary})_1$,’ or by John, in which case it is ‘$*(\text{Mary})_2$,’ with ‘her*’ as its immediate antecedent. Naturally, ‘her*’ is ‘$*(\text{Mary})_1$.’

Our notation is also adequate to dissolve another type of ambiguity proper to linguistic verbs. For example,

(27) John told Paul that he* had been made a full professor.[18]

In this case ‘he*’ is ambiguous between a use of ‘I’ or a use of ‘you’ by John. These interpretations come, respectively, to be:

(28) John told Paul that $*(\text{John})_1$ had been made a full professor,

and

(29) John told Paul that $*(\text{Paul})_1$ had been made a full professor.

For simplicity of the formulation of general principle (P.A) below we shall consider every statement a person X makes or believes to be true, or merely entertains, as subordinated to a tacit strict oo-prefix of the form ‘I think here now that.’ This conforms with the egocentric characteristic of *properties* above discussed.

The preceding examples suggest that the general principle in operation is the following:

(P*.P) Let $\phi(*a_n)$ be an incomplete or syncategorematic clause such that: (i) it contains an occurrence O of $*a_n$, (ii) it does not contain the antecedent A of O, (iii) it is immediately subordinated to the oo-prefix P_n represented by ‘$L[b, \ldots]$’, (iv) P_n is immediately subordinated to the oo-prefix P_{n-1} of the form ‘$L[c, \ldots]$’, and (v) $n > 1$. Let $\phi(*a_n//*_{n-1})$ be the syncategorematic sentence that results from $\phi(*a_n)$ by replacing: (vi) *all* occurrences Ox of $*a_n$ in $\phi(*a_n)$ whose antecedent is A, and which are subordinated to exactly the same oo-prefixes to which occurrence O of $*a_n$ is subordinated, by occurrences of the description function ‘the only thing which is $*\ _{n-1}$’; and (vii) all occurrences of $*a_m$ in $\phi(*a_n)$ whose antecedent is an occurrence Ox mentioned in (vi) by an occurrence of the description function ‘the only thing which is $*\ _m$.’ Then a statement formulated by a sentence of the form

$$— L[\ldots]\ (\phi(*a_n)) —$$

is the same as the statement formulated by the corresponding sentence of the form

(there are *properties* μ_b-ness and $\mu_{b/c1}$-ness such that: — (the only thing which is $\mu_{b/c1} = *a_{n-1}$ & $L[b, \ldots]\ (\phi(*a_n\ //\ *\ \mu_{n-1})))$ —,

provided that the schematic letter or variable ‘μ’ does not occur at all in the former sentence.

In general, I submit that

(Q.12) By $n-1$ applications of principle (P*.P), we may analyze perspectivally every occurrence of a quasi-indicator $*a_n$ in a sentence $\phi(*a_n)$ in terms of the quasi-indicator $*a_1$.

By dropping the suffixes ‘b’ and ‘b/c’ of the predicate variables in (P*.P) we obtain a formulation of the non-perspectival principle (P*.N), and

(Q.13) By $n-1$ applications of (P*.N) we may analyze non-perspectivally every occurrence of a quasi-indicator $*a_n$ in a sentence $\phi(*a_n)$ in terms of the quasi-indicator $*a_1$.

Wayne State University

Received April 4, 1966

---

[18] I owe this example to P. T. Geach, and with it the realization that it exhibits a peculiar kind of ambiguity.

American Philosophical Quarterly
Volume 4, Number 2, April 1967

II. THE GENETIC FALLACY REVISITED

NORWOOD RUSSELL HANSON*

> "If you would know how things are, consider their origins."
> *Aristotle*

I

SOME philosophers of science regard history of science as irrelevant to their professional tasks. Others think history of science essential for philosophizing about science. "Logical Reconstructionists" (e.g., Reichenbach, Carnap, and Tarski)[1] dominate the former group. Individuals of more expansive outlook—Mach, Broad, and Toulmin[2]—adopt the latter position. They feel taunted by the former individuals who say that philosophical reliance on historical facts constitutes a Genetic Fallacy. My sympathies, since they align with the Mach-Toulmin attitude, render me vulnerable to this imputation of fallacy. My wish here is to determine what the Genetic Fallacy is; what is wrong with arguments which proceed directly from facts about *S*'s origin to conclusions concerning "the philosophy of *S*"; which kinds of reflections produce Genetic Fallacies and which constitute acceptable Genetic Explanations; whether every philosophical reliance on historical facts must be an example of this fallacy; and, finally, what the history of philosophy has to do with "doing philosophy." Is there some general conceptual fault underlying the surface features of all "genetic accounts," such that even historians of philosophy should feel anxious about how they express their conclusions?

First, then, what is a *Genetic Fallacy*? Professor Ledger Wood describes it as an "*application of the genetic method resulting in the depreciatory appraisal of the product of an historical or evolutionary process because of its lowly origin.*"[3] Noblemen once scoffed at peasant art. Darwin's Theory was degraded as the product of a "hairless ape." (Would one regard a theorem in mathematics less praiseworthy because its discoverer was uneducated, or otherwise disreputably credentialed?) But not all delineations of the Fallacy are of this form. That is reassuring since, on Wood's account, only a boor would commit the Genetic mistake. Falling into the Fallacy would be, on Wood's view, to be a certain kind of person. But there must be less *ad hominem* versions of anything properly called "a Fallacy."

It has been urged that a Genetic Fallacy occurs *when an explanation of S confuses the temporal order of actual events leading up to S with the logical order in which S must be generated.* Thus the final, elegant, stepwise presentation of a scientific argument is usually the last thing to be associated with that piece of reasoning. The initial groping toward, for example, the Law of Gravitation differed from the economical demonstrations developed later by Analytical Mechanicians like Lagrange, Laplace, and Poincaré. Anyone who now presented the Law of Gravitation (within Celestial Mechanics) move-by-move as in the original exposition of 1687[4]—on the grounds that it could not be fully understood in any other way—would thereby have committed the Genetic Fallacy as delineated here. (At least the meaning of "fully understood" would be up for examination). Or, let '*S*' represent the Pythagorean Theorem; we should all wish to demarcate the minimal, terse presentation of *S* within modern geometry from the centuries of stumbling search which led to a first viable formulation and proof of S (2600 years ago). To suppose that the actual episodes leading up to this discovery must again be fully traversed in thought by anyone who would comprehend its content and purport—this would be to commit the Fallacy as interpreted in this paragraph. The historical progression to S must always be distinguishable from the logical progression to S. The former consists in contingent steps terminating in S (an event, or a claim); the latter consists in "necessary" steps within any progression (from event to event, or from step to step in an argument) which can secure that S.

---

* My gratitude is extended here to Messrs. R. S. Brumbaugh, C. B. Daniels, R. J. Fogelin, T. A. Goudge, M. C. Nahm, J. B. Smith, and R. A. Smyth—all of whom made suggestions which improved this paper beyond what it once was.

[1] As examples of such work see H. Reichenbach, *Elements of Symbolic Logic* (New York, 1947) and *Philosophical Foundations of Quantum Mechanics* (Berkeley and Los Angeles, 1944); A. Tarski, *Introduction to Logic and Methodology* (Oxford, 1941); Suppes, Sugar, and McKinsey, "Axiomatic Foundations of Classical Particle Mechanics," *Journal of Rational Mechanics and Analysis*, vol. 2 (1953), pp. 253–272.

[2] See for example, E. Mach, *Principles of Physical Optics* (London, 1926), and *The Science of Mechanics* (Chicago, 1893); C. D. Broad *Perception, Physics and Reality* (Cambridge, 1913); S. E. T. Toulmin, *Philosophy of Science* (London, 1953) with J. Goodfield, *The Fabric of the Heavens* (New York, 1961), and *The Architecture of Matter* (New York, 1962); M. Hesse, *Forces and Fields* (Edinburgh, 1961).

[3] In D. Runes, *A Dictionary of Philosophy* (New York, 1942).

[4] I. Newton, *Philosophiae Naturalis Principia Mathematica* (London, 1687).

Another presentation is this one: The *Genetic Fallacy obtains if an account of how S came into existence is confounded with an account of the structural constitution of S, here and now.* The historical progression is now confused not with a logical progression *up to S* (as in the last paragraph), but rather with the conceptual analysis of what *S is.* Genesis-talk is here substituted for an account of "the essence of S," rather than for "argument supporting S," as earlier. An engineer's account of the present deployment of stresses, strains, and lateral forces within a steel-framed, glass paneled building need not proceed *via* an account of the original planning of the structure, nor the builder's activities while constructing that edifice. Nor need the engineer delineate procedures necessary for erecting *any* such building—not if he aims to analyze the stresses and forces within this particular structure.[5] Similarly, the formal validity of a provable proposition within Elementary Number Theory is independent of any sketch concerning the genesis of that proposition *qua* the terminus of an argument traversed through time.[6]

Some urge that a Genetic Fallacy obtains *when a proposed explanation of S treats it as a "disguised form" of some earlier state in a sequence (historical or logical) generating S—ignoring thereby the unique peculiarities S may have in itself.* Thus, parodies of mature men as being nothing more than outsize infants (a ploy of popular psychoanalysis), or of thinking as being nothing but "silent talking" (popular epistemology), or of symphonic music as being only sophisticated and refined "tribal noises" (popular anthropology), or of mathematical research into many-dimensional spaces as being but a precious adjustment of hominids to their biological environment—these "nothing but" reductions, although illuminating, make it difficult to appreciate such *designata* for what they are *now*, so much stress having been given to the characterization of maturity, thinking, and music-making as being but the contemporary attention-getting stages in an ancient series of events. Granted, the Empire State Building does have affinities with the Nomad's tent; the Queen Mary resembles the Argonaut's raftlike vessel in some ways; our national society is in many ways Neanderthloid . . . but such pronouncements can only be analogical preliminaries to real analysis and understanding. These latter require delineating differences between S and its genetic antecedents, as well as noting the similarities.[7] Anyone who substitutes such headline allusions for the analysis itself thereby commits this form of the Genetic Fallacy.

The Genetic Fallacy is sometimes said to arise *when a trivial necessary condition of some highly developed state of affairs is presented as if it were a sufficient condition for that state.* Thus, we noted that mathematicians are men, and that men often behave as they do for reasons that are biologically inspired. From these considerations some suggest that mathematical theories are but one more effect of human activity, not different-in-principle from ingestion and respiration. In this frame of mind love is analytically reduced to reproduction, charitable activity becomes characterized as disguised self-interest, scholarly debate is a jungle encounter-to-the-death between savage combatants . . . etc. But such things as love, charity, and thoughtful argumentation can be described in analytical terms[8]—as meaningful, composite *denotata* here and now. True, perhaps, these *denotata* would not even exist were there no biologically oriented hominids of the sapiential species. But that is no explanation of the *specific* features of mathematical proof, of the intricate emotions of love and the sentiments of charity, and of the conceptual structure of intellectual exchanges. The real significance of Gödel's theorem does not rest in its having been produced by an animal.

---

[5] Think of the professional engineer lecturing on the "formal" features of a flying buttress, or a Gothic arch, or a high transept, such as one finds in the cathedrals of Amiens or Chartres. He will not give the local building histories of such structures. Nor need he set out the features of buttresses, arches, and transepts in general. He will, rather, give an analysis of these particular architectural specimens *as they are*, not as they came to be, and not as all such specimens should be. Compare Aristotle: "For the definition of house-building includes and presupposes that of the house; but the definition of the house does not include nor presuppose that of house-building; and the same is true of all other productions" (*De Partibus Animalium*, 646a).

[6] Johnny's teacher may have had to stand on her head to get him to see that $\sqrt{4} = +2$ or -2. But such persuasive pedagogy is independent of the validity of the claim that $\sqrt{4} = +2$ or -2.

[7] The "lumping" and the "splitting" behavior, so well known to taxonomists, must be kept in balance by philosophers—especially while invoking genetic references. Lumping without splitting is empty; splitting without lumping is blind.

Yet another, somewhat blatant, form of the Genetic Fallacy *consists in evaluating an argument by references (defamatory or congratulatory) to the "psychological type" of the arguer.* Critics of the methodological shortcomings of psychoanalytic theory have occasionally found their own criticisms attacked as being but manifestations of "father hostility," or a "wish fulfillment," or perhaps even a "death-wish"; as if these further observations had relevance to the intra-theoretic issue. Even if true, they are beside the point. Imagine a geometer smarting from a critic's claim that inconsistency ruins his new geometry. The wounded scholar retorts that this is only to be expected from a former student (the critic) who fared badly in a course in Conics once given by the geometer himself. True or false, this in no way affects the inconsistency-issue; to suppose that it might is to commit another form of the Genetic Fallacy—an ungraciously *ad hominem* form.

Even more pernicious variations of the Fallacy *consist in enlisting as support for a claim S all those intra-cranial episodes leading up to the enunciation of S.* Newton's cerebrations while puzzling out the Law of Gravitation *may* have been identical with our terse demonstrations of that law. But this *de facto* identity is irrelevant, however enlightening it may be about Newton's inner life and times.[9] Kepler's *De Motibus Stellae Martis* faithfully records his every insight, bewilderment, and blunder. It is thus more informative about him, and about Celestial Mechanics in the 17th century, than about the cogency of his arguments terminating in the claim: "Mars's orbit is an ellipse one of whose foci locates the sun." Of course, information about Kepler and 17th century science may be just what we want. But not in answer to the questions: "Was Kepler's argument redundant in any way?; logically inconsistent in any way?; mathematically inelegant in any way?" Similarly the question "Was Newton's reasoning leading to the Law of Gravitation formally sound?" concerns not why certain calculational techniques and hypotheses appealed to him; nor is it a question about Newton's private mental history. (What would it matter that he entertained purple spots as visual sense-data whenever he thought about gravitation?) Rather, such a question concerns whether Sir Isaac's argument, considered as inferentially-linked propositions expressed by sentences and formulae within the *Principia*, is (timelessly), a *good* argument. Some of his arguments *were* good ones. A few were not so good. But such judgments become possible for any philosopher who knows what Newton's arguments *were*, how they were structured, and how they are logically vulnerable. Such appraisals do not require further insight into Newton's psyche.

Two fallacies from the same basket are The Intentional Fallacy and The Effective Fallacy (or "Affective Fallacy").

If one's appraisal of a work of art were shaped by information about the artist's intentions when creating it, that would exemplify the Intentional Fallacy. Myron's aim in sculpting *Discobolus*, Beethoven's objectives in writing *Eroica*, Picasso's goal in painting *Guernica*—these are beside the point of evaluating such works of art as they are, *via* objective criteria other than the artist's motivation. (Indeed, the artist himself may be in the worst position for appraising his own work.) Objective analysis of a creation is logically different from psychoanalysis of its creator.

The obverse lapse occurs in the Effective Fallacy wherein *the beholder of a work of art (or of an argument) confuses its emotional or psychological impact with its objective merits.* But again, the felt effect of an argument is independent of its conceptual structure; as the latter is appraisable apart from the subcutaneous intentions of the arguer, so also can it be evaluated apart from the pancreatic responses of the listener. Intentions and effects never illuminate the form and structure of the thing itself, whether it be a work of art like a painting, or a work of thought, like an argument.

All these versions of the Genetic Fallacy, then, consist in judging the formal acceptability of an argument. But that judgment is drawn *via* reasoning within which there figures (covertly) an inference-linkage of the form "when historical facts α, β, γ . . . obtain, then the ensuing argument is inferentially sound"; or "given a psychological sequence, *à la* α, β, γ, then the resultant argument is conceptually sound"; or "since the speaker's intentions were α, β, γ, then his argument must have formal features A, B, C"; or (finally) "since my felt reactions to the argument are α, β, γ, the argument must have logical features, A, B, C." Sundry further variations are possible, of course.

---

[8] By which it is not meant, of course, that all claims about love, charity, and thought are analytic. Many analyses of structures, e.g., bridges, societies, arguments, are constituted throughout of contingent propositions.

[9] Of this more in a moment.

Beyond this similarity these several versions of the *Genetic Fallacy* apparently have very little in common. Beyond consisting in reasoning-chains at least one of whose premisses is contingent (i.e., historical, psychological, factually descriptive . . . etc.) and the conclusions of which are all judgments evaluating formal features of some structure, there seems barely more than a family resemblance between them all. But they do perhaps signal *some* conceptual impropriety in attempting to support the "rightness" of an argument by reference to facts leading up to the "discovery" of such an argument. The imputation behind all disclosures of the Fallacy is that a recitation of facts leading up to (or away from) an argument, concept, or conclusion, is irrelevant to appraising the structure of such an argument, concept, or conclusion.

II

Now, is it really true that the structure of S and the genesis of S are completely unconnected? Must they be unconnected?

The answer is "no," by way of a consideration already hinted at. There *could* be an identity between the steps needed to delineate the structure of some S, and the steps actually traversed in generating S in the first place.[10] Why could not Pythagoras initially have set out his theorem, and its proof, in the same elegant, terse mode of exposition we now use to demonstrate its validity? There are good historical reasons why not, but no logical ones. And why could not Newton have moved along the inferential steps (through his argument which terminated in an articulation of the Law of Gravitation), the *same* steps we use today in Analytical Mechanics succinctly to demonstrate that very Law? We considered this possibility earlier and rejected it as irrelevant. Were we hasty?

It *could* help our understanding of S's structure for us to know about the genesis of that S. Thus an early genus of elephant (*Moeritherium*) lacked a trunk; it had a ponderous lower jaw with protruding, shovel-like incisors. These were probably used to scrape for food lodged in the ground. Such incisors worked against a horny pad covering a fleshy "Cyranoid" extension of the face. This was the primitive forerunner of the trunk. But the shoveled food could not easily be conveyed to the beast's mouth by its lips; that task was taken over by this facial extension—this proboscoid pad. The (paleozoologically) probable result was that this new organ became ever more efficient as an instrument for gathering food, independently of the lower jaw and incisors. A progressive shortening of the lower jaw followed in some species, as well as a dimunition and even the loss of the lower incisors. The trunk of our modern elephant was gradually generated.

Here, then, is a genetic account of how the elephant's remarkable trunk is insofar *understood*; its structure (as a boneless, tough-skinned musculature), is in some way comprehended by an account like this one—an unashamedly genetic account.[11] It seems at least plausible that structural questions in disciplines like paleobotany, paleontology, archaeology, historical sociology, geology, etc., might best be set out in genetic form. Is there some logical reason why this should not be the case? Aristotle nowhere provides one. Nor do his successors. The point is that for some X's, learning of the development of X can convey knowledge of the structure of X here and now.

Still, even within these "historical sciences," where genetic explanations do seem to have a role, impropriety still marks anything grossly analogous to a Genetic Fallacy. Given the diversity of kinds of argument which we have seen to count as Genetic Fallacies, it is small wonder that no *formal* features of this fallacy are ever delineated. Fallacies such as "denying the antecedent," "asserting the consequent," "undistributed middle," and "illicit process," etc., have formal features which we can identify as constituting the improprieties in question. When the variables in the formal schematism are evaluated (and psychologically disguised) the entire inference constitutes a fallacy. Textbook fallacies consist in both formal and psychological components. The formal constituent consists in the abrogation of some rule of deduction: this is the logical invalidity. The psychological component consists in that invalidity being cloaked in an apparently cogent argument; the error must go undetected by at least one hearer. For, if neither the speaker nor any of his audience is "taken in" by a fallacy (a situation that often obtains in logic seminars), then the proferred inference is merely being named, e.g., as "the fallacy of asserting the consequent"—it is not being appraised as such by one of the group in an attempt to instruct the others. A genuine fallacy, then, has an objectively invalid form, while being construed by someone as being quite valid. A chain of assertions which tricks no one is not a fallacy for anyone.

---

[10] This may contradict Aristotle, who argues: "Now the order of actual development and the order of logical existence are always the inverse of each other" (*loc. cit.*).

[11] Cf. J. Maynard Smith, *The Theory of Evolution* (Harmondsworth, 1957), pp. 245–256. I owe this reference to an article by T. A. Goudge called "The Genetic Fallacy" which appeared in *Synthese*, vol. 3 (1961), p. 44.

In obvious cases we all feel that something of a general nature has gone wrong in a Genetic Fallacy. A man who argues for the formal validity of some argument or conclusion *may* take us through all the steps leading up to that conclusion, the very steps indeed which he actually traversed in puzzling out the proof originally. "Einstein teaching" *may* have been just an audible repetition of "Einstein thinking." Why would we not reject *this* as a Genetic Fallacy? What is it in Genetic Fallacies that is analogous to the deductive breach we signal in the Formal Fallacies? What, in short, is *formally* incorrect about arguing from a delineation of historical facts to a conclusion concerning analytical structure? What is wrong with arguing from "Pythagoras realized that P in 600 B.C.," and "Pythagoras perceived that *if* P *then* Q in 599 B.C." to the conclusion 'Q'? (where 'Q' represents, e.g., "All equiangular triangles are equilateral.") In fact, nothing is wrong with this inference. It is not an example of a Genetic Fallacy, even though the premisses are dated (and hence contingent in some sense) while the conclusion is not dated (and moreover, is possibly necessary).

This way of putting matters suggests a line of inquiry. One would not prove the validity of the Pythagorean theorem by tracing a right angle on sheet metal, constructing squares on each of the legs, and then cutting out and weighing the square constructed on the hypotenuse *against* the combined weight of the squares constructed on the other two legs—as by placing the metal plates in opposite pans of a beam balance. (Does the hypotenuse-square just balance two leg-squares?) This, obviously, could not prove or disprove the Pythagorean theorem. Why not? What would be the formal error in thinking otherwise? The error isn't formal at all; it is not something concerning only how propositions have been combined, unpacked, and decomposed into less general claims. It does not concern the form of argument at all. Perhaps the issue is a larger one—involving distinctions between logical types, or at least between different levels of discourse.

III

Notice that historical details concerning the genesis of S are always factual in nature (at any rate in large part). What else could historical details be? (There are no necessary historical facts.) Historical claims are expressed in consistent, synthetic propositions (i.e., the negations of which are consistent), which are *contingent* (i.e., in different circumstances they need not have been true—they are vulnerable, defeasible claims), and *a posteriori* (only experience provides the knowledge required veridically to generate such an account). But, is an analysis of the structure of a proposition, or of a concept, or of an argument factual, synthetic, contingent, and *a posteriori*? To have proved the validity of Pythagoras' theorem within Euclidean geometry is to have done something for all time. Further observations would not be to the point; tomorrow's facts cannot upset the validity of today's well-made formal analyses; new circumstances of inquiry would not change one mathematical, logical, or semantical conclusion—if drawn properly to begin with. New experiences will never qualify the rectitude of one's geometric determinations. And, most importantly, the intra-geometrical acceptability of the Theorem in question is announced in a proposition which is not synthetic. The claims "the Pythagorean Theorem is deducible within Euclidean Geometry" and "all equiangular triangles are equilateral" do not have negations which are synthetic in any straightforward way. There are good reasons, indeed, for thinking those negations to be demonstrably inconsistent. The full meaning of "Euclidean Geometry" involves the Pythagorean Theorem as one of its understood constituents: it requires equiangular triangles to be equilateral. A geometry without these theorems is just not Euclidean. So the negation of such claims does not state what is factually false; it raises questions about whether the hearer really comprehended the meaning of the terms involved at all. The negations of these claims are conceptually untenable. And if the negation of S is not synthetic, S is not synthetic either. Therefore "The Pythagorean Theorem is deducible within Euclidean geometry" is not synthetic.

What about the zoologist's physiological analysis of the structure of the elephant's trunk? Surely that will consist throughout of factual, synthetic, contingent, and *a posteriori* claims—premisses and conclusions both. Granted, not all analyses of an *S* are set out in analytic propositions about *S*; indeed, most are not.[12] Still, these analyses, though wholly contingent, will not be claims functioning on the same logical level as would assertions concerning the *de facto* historical evolution of the elephant's trunk through time. There are deep methodological differences between the way in which modern physiologists justify their "analytical remarks" about the structure of the elephant's trunk (its musculature, its neural network, its epithelial layers, etc.), and the way in which a paleozoologist would justify his claims about what stage of development the elephant's trunk must have been in two million years ago. Different from both types of (contingent) remarks would be the veterinarian's claims about the (special) structure of this particular elephant's trunk (Jumbo's); finally, statements concerned with oddities in the development of Jumbo's trunk through time (its shape and musculature at one year of age, its tough epidermal casing at five years, etc.), these factual remarks would be different again from all the others. So even within the large class of propositions designated by "factual, contingent, synthetic, and *a posteriori*" there will still be myriad levels of language, transitions between which are not easily effected. Certainly when such transitions *are* undertaken, the slide should be explicit and self-conscious with no lack of clarity about inferences which may be slipping from premisses within one language-level to conclusions from quite a different level. For example, an analysis of Jumbo's trunk will warrant conclusions about *all* elephantine trunks only on the basis of some carefully formulated assumptions. Questions of the validity of such inferences should be addressed case-by-case. Some may be quite illegitimate; others will appear acceptable only when floated upon some "buoyant" presuppositions. Political claims purport to be "factual, synthetic, contingent, and *a posteriori*," but so do most of the claims of theoretical physicists. When politicians instruct scientists what to say such sub-languages interpenetrate; but the impropriety is grotesquely apparent.[13] Distinct types of factual discourse rarely mix, especially within single arguments. The timely detection that they do not mix in a particular case is sometimes designated as the detection of a fallacy. The paradigm *non-sequitur* consists in an injection of the wrong kind of factual claim within an argument constituted wholly of another kind of factual claims.

Perhaps Genetic Fallacies are somehow like this? Perhaps they are apparently valid arguments which, without explicitly signaling it, proceed from premisses at one level of discourse (e.g., historical facts) to a conclusion of quite a different kind (e.g., an analysis of some contemporary structure, like a locomotive, or a lemma). But, perhaps noting only that Genetic Fallacies, diverse as they may be, all consist in sliding inferentially across levels of discourse—perhaps this is not to have noted enough. (This much alone would make Genetic Fallacies nothing more than a species of *non-sequitur* argumentation; they are more.) One might persuade himself to conclude that Pythagoras stated his Theorem in 599 B.C. in a form restricted to "3–4–5" right triangles—this factual conclusion seeming to follow from premisses which abstractly set out the Theorem in the terse, elegant, and general form we know it in now. Such an argument would proceed across types, from non-factual premisses to a factual conclusion. But *that* doesn't mark the argument as a genetic fallacy since it *terminates* in a reference to genesis, not *vice versa*. That is, inferring across levels of discourse may be a necessary condition for any argument to be genetically fallacious, but it is not sufficient. The slide must also be from premisses which are relatively factual, contingent, and historical, to a conclusion (at a different language-level) which is relatively analytical and "non-historical." The reverse pattern of inference (analytical premisses →historical conclusions) may be objectionable, but it is not the pattern of a Genetic Fallacy.

---

[12] Analyses of force-networks in arches, bridges, aircraft, and boats are uniformly constituted of empirical statements. Yet they are highly general, i.e., non-specific, and in no important way are they conclusions drawn from a set of merely historical factual claims about the local stories of particular arches (". . . de Triomphe"), bridges ("Waterloo"), aircraft ("Enola Gay"), or boats ("Challenger").

[13] Separate languages were forced to coalesce in the infamous Lysenko incident, in the Nazi condemnation of Einstein's work as "non-German," in the Scopes' evolution trial, in the Church's examination of Galileo and Copernicus, and in other sad and sobering cases.

Whenever a conclusion S is "non-factual, analytic, necessary, and *a priori*," premisses concerning the factual genesis of S are obviously of a different logical type from such a conclusion. This is the "Pythagorean paradigm"—an "L-true" proposition which came to be discovered, or articulated, within a contingent context of inquiry.[14] But even were S of the *same* logical type as the steps set out in a genetic-historical account of S,[15] there may still be profound conceptual distinctions to be drawn between the types of discourse forced so to mix.

I have avoided saying that arguments from propositions of one logical type to conclusions of another logical type constitute a formal fallacy. Indeed, I deny just this. The logical type of a proposition rests not on purely formal determinations in any case. Formally the inference from "$2+2=4$" and "if $2+2=4$ then the moon is made of plaster" deductively guarantees the conclusion "the moon is made of plaster." In addition to the uneasiness felt when confronting such a "degenerate" inference, the conjunction across logical types also jars one's sensibilities. Nonetheless, this is as good an instantiation of *modus ponendo ponens* as could be found; the inference is legitimate and invulnerable.

Designating propositions as different in logical type may be more a "metaphysical" move than the result of strict sign-design analysis. Since the proposition-chain above (involving numbers and plaster) is formally acceptable, one's reasons for classifying mathematical claims as being of one type, while factual-descriptive propositions are of quite another—these reasons must be *non-formal* in character. They have nothing to do with obeying, or breaking, the rules of some symbolic, inferential game. Two philosophers, even after agreeing about the inferential rules governing a statement like "$2+2=4$," may still disagree about the proper philosophical characterizations of that statement.[16] Statements of historical fact always operate on a level different from that of statements directed to the analysis of structures (be they arches or arguments). The difference may not always be as dramatic and obvious as where we contrast the mathematical analysis of a theorem (e.g., Pythagoras') with the historical development of a proo of that theorem; the former will be couched in a cluster of analytical (and analytic!) propositions, while the latter must consist in synthetic claims. But the differences between *levels* of discourse remains apparent even when all the claims considered are synthetic, and even when these differences turn on nothing more than a recognition that historical accounts always concern events in the past, whereas analyses seem to be time-independent (or at least they concern the present primarily). It would require much time and effort to set out all the steps John Austin traversed in coming to see that the assertion "I know that α" entails "It is true that α." Still, this latter insight fills a different logical space from that occupied by any history of Austin's reflections thereupon. *That* X entails Y, fills a different logical space from *coming to see that* X entails Y.

In several places I have discussed arguments which proceed across levels of discourse.[17] Again, let me urge here only that the "fallacious" character of genetic accounts which purport to illuminate "structures" (e.g., arches, aortas, arguments) consists in just such an arguing across the language levels, despite the initial and terminal claims being of the same logical type. Within the class of contingent claims the argument may proceed from relatively "historical" premisses to relatively "analytical" conclusions. Classes of propositions homogeneous as to logical type are thus internally striated with myriad micro-levels of language. "Apples contain sugar" is contingent;

---

[14] Of course there are many valid deductions of necessary conclusions from contingent premisses. "My father is male" (a necessary claim) follows from the premisses (1) "My father is bald" (contingent) and (2) "All bald people are male" (also contingent). So there is no general logical impropriety against inferring across types. But genetic fallacies are a peculiar sort of cross-type inference. It is this special peculiarity which is singled out for attention when the imputation of "genetic fallacy" is advanced. Cf. N. R. Hanson, "A Budget of Cross-Type Inferences," *The Journal of Philosophy*, vol. 58 (1961), pp. 449–470.

[15] Thus it is with an engineer's analysis of the stresses operative in Waterloo bridge, as against the historical account of how that bridge came to be built at all. Both the analysis and the account will be factual, synthetic, contingent, and *a posteriori*. Still, to infer from this that there is no difference between the language-levels of the analysis and of the history would be a mistake.

[16] Think of J. S. Mill and D. Hilbert disputing over the logical type of $2 + 2 = 4$, where no formal demonstration would reduce them to agreement.

[17] N. R. Hanson, "Imagining the Impossible," *Analysis*, vol. 18 (1958), pp. 86–92; "Its Actual So Its Possible," *Philosophical Studies*, vol. 10 (1959), pp. 69–80; "A Budget of Cross-Type Inferences," (*op. cit.*).

but so is "2 apples *plus* (i.e., "simply combined with") 2 apples results in 4 apples." These claims function at different levels of contingent discourse—and any argument which purports to bind them as premisses and conclusion must therefore be critically scrutinized. The same is true of these statements: "This apple fell off a tree planted by Sir Isaac Newton" and "This apple is 80 per cent $C_6H_{12}O_6$ by dry weight." Any argument which would make this chemical analysis dependent upon the ancient history of this particular apple would similarly require strict scrutiny.

This may explain the diversity one finds in those claim-clusters we call "Genetic Fallacies," nothing being common to them all save that they are so-called. And one further feature is shared by them; they all constitute inferences from historical descriptions, to analyses which, if sound at all, are so because of considerations which demarcate them from anything "*merely* historical." That this apple is 80 per cent sucrose is established independently of data concerning its past history.

Some analyses of structures may differ so much from what one encounters in formal contexts, may be so "developmental" indeed, that the issue might seem to swing toward the opposite pole of resolution. How could there be any satisfactory analysis of the administrative structure of NASA without an historical delineation of the events leading up to Sputnik I in 1957, and Luna 9 in 1966, and the repercussions to those events all over the world? Similarly with the present "parliamentary structure" of our Congress, most "analytically" presented in courses on Government. Could such an administrative anatomy and hierarchy be intelligible without some evolutionary account of why and how our present political and governmental institutions came to be interrelated as they are? Like the French Revolution and the collapse of the Third Reich, such *denotata* seem *essentially* to require genetic explanation. These examples, though less offensive (as concerns cross-type inference) than our contrived Pythagorean example, seem yet to involve switches-in-discourse, from genesis and history to structure and tables of organization—from facts about the past to analyses of present complexes.

Perhaps the major message before us devolves into the near-tautology that *differences must be distinguished*; matters of genesis must be distinguished from matters of analysis. Never should the two be set up as endposts holding down one and the same argument. This, even when the genetic account and the structural analysis are constituted of claims of the same logical type—i.e., even when they are all "factual, synthetic, contingent, and *a posteriori*." Descriptions of Jefferson's childhood differ from descriptions of the form of arguments within the Declaration of Independence. Yet they are all *descriptions*—between which, however, inferences should be conducted only with considerable caution, and with explicit attention being given to the inferential links which may be felt to bind the premissory descriptions to the conclusion-descriptions.

Back to Newton; consider again the hypothetical situation wherein every step in his original "thinking out" of *S* might be identical with our present day expositions of *S*. It is logically possible that such a thing could happen. But even if it did, a distinction could still be drawn. For every mental step Newton makes has an implicit date, and a time, attachable to it. It will always be true or false to say "Newton made it to *S-minus-4* at noon on June 1, 1686." However, when we in our terse expositions of *S* rest momentarily at *S-minus-4* there is never, not even implicitly, a date and time attached. This "datability" always makes a genetic account a factual one. Whereas the obverse situation suggests that an analysis of *S* (even if "factual") will be on a different level of discourse from the history. A geologist's analysis of the strata beneath us, one that might be useful to an architect or an engineer, will not be at all like the historically episodic account given (of those same strata) by a paleogeologist. Era-terms like "Cambrian" and "Devonian" will be essential to the latter, but irrelevant to the former (save as mere names).

Giving an analysis of *S* (here and now), then, requires a different manner of discourse from that involved in giving an evolutionary description of *S*'s development out of *S-minus-n*. These linguistic tasks have different objectives, employ different techniques of exposition and argument, and are subject to different criteria of adequacy. This, despite the fact that both the structural analysis, and the evolutionary description, may be set out in discourse of the same logical type, i.e., in "factual, synthetic, contingent, and *a posteriori* propositions." (Even though the difference here is not as apparent as when one contrasts the conceptual analysis with the *de facto* history of the Pythagorean Theorem, it remains difficult to resist concluding that in *every* case of the Genetic Fallacy some confounding of levels of discourse is involved. This, although premisses and conclusion are of the *same* logical type.)

IV

A residual problem remains. For what are the interconnections between (say) History of Science and Philosophy of Science—between the historical development of scientific arguments and the philosophical analysis of those arguments? How does an etymological account of the genesis of some scientific term (e.g., "force" or "electron"), relate to a semantical analysis of the meaning of that term, as used in contemporary reasoning? And what then, derivatively, are the connections between the History of Philosophy *and* Philosophy; between the facts about past philosophers (and their philosophies) *and* the doing of philosophy—this latter being a notoriously non-factual undertaking. Have we here but two more pale reflections of the Genetic Fallacy? For, surely, philosophers often wax historical and scholarly in the course of their analytical disputations.

How can data from the history of thought, or from the history of science, help conceptual analyses here and now?

We must all feel uneasy when experimental psychologists, social anthropologists, and "kultur vultures" sail too glibly into strictly logical discussions, those concerned with the semantical content of technical concepts and the structure of complex arguments. How satisfying sometimes to hear the rebuff: "that is just a matter of fact," or sometimes, "that is merely a Genetic Fallacy!" Throughout the History of Philosophy illumination has often been scattered within clouds of conceptually irrelevant historical detail. Often a question is asked (perhaps about whether a given conclusion follows within some argument) and, too predictably, the air becomes charged with quotations from Plato and Aristotle "who thought that it did follow," and from Spinoza and Gassendi "who thought that it did not." Indeed, a signal irresponsibility among philosophers is to serve up a tray of facts when what is needed is the scalpel of analysis. Philosophers of Science know too well how scholars sometimes dull the cutting tools of thought by thrusting them into the historical gravel. Conceptual clarity is the result of uncluttered analysis; allusions to actual occurrences in the History of Science can be *at most* illustrations of arguments which commend themselves, if at all, on rational grounds alone.

That *S* is done universally does not make *S* the universally correct thing to do. Fifty thousand Frenchman can be wrong. (Fifty million can be!) That all past and present scientists *do S* or *say that S*—or are reported by historians to have done or to have said that *S*—this, again, does not in itself make *S* the thing to do or to say. Philosophy of Science (as with all philosophy) is never just a recollection, rehearsal, and recitation of what *is* done and said, or of what *has been* done or said. It must also be an analysis and an appraisal of the rationale and justification of scientists for doing and saying what they do. Appraisals, interpretations, and judgments operate within levels of language quite distinct from those within which factual, descriptive, chronicloid treatises are composed.

Suppose all this to be true. Grant with me that History of Science and Philosophy of Science do operate within different levels of discourse (the former being concerned with descriptions of facts, the latter with conceptual analysis). Given that the several forms of Genetic Fallacy are alike only in that they intertwine and confuse just these distinct levels of discourse, then it would seem possible that the philosopher of science who depends upon History of Science in his analyses is forever on the brink of the Fallacy. He regularly risks injecting "what is the case" into arguments whose conclusions are of a wholly different nature.

Perhaps one should admit that, in strict logic, the History of Science has no inferential function within Philosophy of Science. The former consists in stating facts, past and present.[18] The latter discipline, however, consists in providing analyses, appraisals, interpretation, and insight. These are never obviously factual undertakings. If what an historian of science says is false it is likely to be factually false. If what a philosopher of science says is false it is not likely just to be factually false. His error will be more one of understanding and of inference than of data and dates. The attempt to purchase the validity of a philosophical argument by reciting historical facts deserves the scolding inevitably to ensue.[19]

Still, some slight softening of this proscription may seem desirable, and for two reasons. Historians and philosophers of science are both concerned with the structure and function of scientific argument. Although philosophy "in general" may have no subject matter, Philosophy of Science certainly has. Broadly, it has the same subject matter as does the History of Science—*science* itself. Without a detailed understanding of the *de facto* development and present state of modern science, the analyses of philosophers of science must be barren and unilluminating.[20] But the other philosophers, those anxious about "historicism," view the heaped-up facts of history as equally unilluminating. Histories of science are too often chaotic and diffused (they seem not to reflect monochromatically). Spectra of concepts and arguments result, because terms like "law," "theory," "verify," "observe," "hypothesis," etc., are used by historians uncritically, whereas philosophers know well that these concepts are crying for analysis. Yet, on the other hand, to the historian of science, *pure* Philosophy of Science (*à la* Carnap, Taski, and Braithwaite) appears equally unilluminating; it does not enlighten any particular thing or problem within science such as would be recognized by scientists-at-work. Little in the record book of physics appears any more comprehensible due merely to the symbolic studies of Axiomatizers and Reconstructionists.[21]

---

[18] Granted, these historical facts may range over diverse levels of abstraction. Of this more later.

[19] Cf. N. R. Hanson, "The Irrelevance of History of Science to Philosophy of Science," *The Journal of Philosophy*, vol. 59 (1962), pp. 574 ff.

For the historian of science, philosophical analyses seem empty; for the philosopher of science, the historian's factual compendia are blind.

The issue is not really complicated. One cannot philosophize about S until he comprehends the facts about S—until he knows what S is. Having such facts at one's command will not by itself guarantee the correctness of an analysis of S. Nonetheless, given that one's analysis is conceptually valid, it will still not seem to be *significant* unless grounded in the historical facts concerning S. The Scylla and Charybdis of this discussion emerge, then, as *Tautology* and *Chronicle*, the former (Philosophy of Science) being analytically valid but insignificant, the latter (History of Science) being factually detailed but formless. Without thoroughgoing interpenetration between these disciplines Philosophy of Science (*sans* history) may be true, but trivial; History of Science (*sans* Philosophy) will be fact-studded, but structureless.

A second reason for smoothing out any abrupt disciplinary division between History of Science and Philosophy of Science is that the former is never a *purely* fact-stating discipline, and the latter is not exclusively analytical. Earlier we considered an inference from

1. Pythagoras realized that P in 600 B.C.

and

2. Pythagoras perceived that if P then Q in 599 B.C.

to

3. Q

There is nothing illegitimate about this inference because the language used to set out the chronological facts in (1) and (2) articulates the historian's deeper commitment to P and to *if-P-then-Q*—from which Q follows (timelessly). To say "Pythagoras realized that P in 600 B.C." is to do *two* things. It is (a) to concede it to be the case that P, and (b) to date Pythagoras' first awareness of that fact. This is thus a "mixed" claim, not a purely contingent matter of historical fact. If challenged as to whether or not it is the case that P (e.g., whether all Euclidean equiangular triangles are equilateral), the historian cannot respond effectively by serving up only more historical facts. The same applies to many other pronouncements typical of historians of logic, of mathematics, and of science.

For example:

In 600 B.C. Pythagoras *proved that P*
In 600 B.C. Pythagoras *established that Q*
In 600 B.C. Pythagoras *saw that R*
In 600 B.C. Pythagoras *demonstrated that S*
In 600 B.C. Pythagoras *revealed an inconsistency* in T
In 600 B.C. Pythagoras *discovered that U*

"In 600 B.C. Pythagoras proved that P" entails that

(a) it is the case that P
(b) there exists a proof that P obtains
(c) Pythagoras discovered that proof in 600 B.C.

---

[20] This is a verdict some physicists have been heard to pronounce on H. Reichenbach's *Philosophical Foundations of Quantum Mechanics*, *op. cit.*; P. Suppes and J. C. C. McKinsey's *Axiomatic Foundations of Classical Mechanics*; and R. B. Braithwaite's *Scientific Explanation* (Cambridge, 1953).

[21] There are such things as mathematical reconstructions, of course, but these differ from what are called "logical reconstructions," both in means and ends. Dirac did much more than merely logically re-structure Quantum Mechanics as he found it.

(c) states an historical fact. (a) and (b) do not. (c) is contingent; in a different world it need not have obtained. (a) and (b) are not contingent, at least not when P takes a value like "All equiangular triangles are equilateral." This does not depend on how things are *in fact*, but only on the internal properties of the Euclidean "space game."

"In 600 B.C. Pythagoras established that Q" entails that

(a) Q
(b) it can be established that Q
(c) Pythagoras did so in 600 B.C.

(c) if known at all, is known *a posteriori*. But, when Q takes a value like "The square on a Euclidean right triangle's hypotenuse equals in area the combined squares constructed on the other two sides," (a) and (b) would not seem to be good candidates for unrelieved *a posteriority.*

"In 600 B.C. Pythagoras saw that R" entails that

(a) R
(b) Pythagoras realized this in 600 B.C.

When R represents "Cubes have twelve edges" (a), unlike (b), is not synthetic in form. Its denial, unlike that of (b), is demonstrably inconsistent.

"In 600 B.C. Pythagoras demonstrated that S" entails that

(a) it is true that S
(b) the truth of S is demonstrable
(c) Pythagoras demonstrated just that in 600 B.C.

(c) is informative as to what contingent possibilities, out of unnumbered alternative possibilities, do in fact, obtain. But if S stands for something like "The internal angles of a Euclidean triangle add up to 180°" there are no alternative possibilities. (a) and (b) are not informative then, not in the sense that (c) is. Rather, (a) and (b) are more like the exfoliage from out of a set of formal premisses—"game-like" in their self-contained tokens, moves, and rules, and internalized as concerns their objectives of derivation and demonstration.

Therein lies the conceptual magic of "mixed" propositions. They are always "molecular"; logical decomposition easily reveals their component parts *seriatim.* Each of these parts is, in its turn, propositional, i.e., that which is expressed by declarative sentences in the indicative mode of speech, and of which "true" or "false" can be predicated. If these constituent parts are also molecular, they are nonetheless of more restricted generality than the "higher-order" claims from which they were unpacked. The theory of deduction ensures that.

But more: The logical types of the component claims constituting these "higher-order" molecular propositions are never all the same. A molecular claim *all* of whose constituent parts were non-factual, necessary, *a priori*, analytic, and uninformative would not be a mixed claim at all. Rather, it would be some possible statement out of formal discourse—out of mathematics or logic. If, on the other hand, all of its constituents were factual, contingent, *a posteriori*, synthetic, and informative, the claim would be "unmixed" in the opposite sense. "In 600 B.C. Pythagoras learned that the earth is a sphere" is not perplexing in any way as to its aggregate logical type.

Are there absolutely atomic statements which could be ultimately constitutive of the "mixed" molecular claims which now hold our attention? The further development of this paper in no way depends on an answer to that question. All we need to note is that those assertions typical of history of science, assertions such as "In 600 B.C. Pythagoras revealed an inconsistency in T," are mixed and molecular. They are molecular in that they can be unpacked into "logically simpler" constituents; they are mixed in that those constitutients are always of at least two different logical types, or language levels.

These examples of reportorial discourse are not unusual within the best work in History of Science. Yet they consist not in purely factual claims. Profound conceptual commitments are packed in at every turn. Hence, from a selective clustering of such statements an historian might legitimately draw a conclusion purely philosophical, i.e., analytical. Such an inference *could* dispense with the anterior references "In 600 B.C. . . .," "In 500 B.C. . . .," etc. But since such calendrical references are the very stuff of history, there is no overriding reason for their elimination, just so long as one remains keenly aware that the analytical[22] conclusion follows only from the *analytical*[22] *parts* of the premisses, and never from the factual parts.

The molecular premisses of history must be "atomized" before philosophical conclusions can be drawn from them transparently.

One must conclude then that History of Science

---

[22] "Analytical" in the broad sense intended—as in "analytical mechanics" and "analytical chemistry"; *not* in the restricted sense of "analytic proposition."

(if conceived of as a purely fact-stating discipline) and Philosophy of Science (if conceived of as a purely analytical discipline) are not logically related at all. Inferentially they cannot be bridged. Premisses from the one can never yield conclusions in the other. Demonstrations cannot be extracted from dates. To claim otherwise would be to commit some gross form of our capacious and elastic Genetic Fallacy.

But the risk of inferring from this that there is no connection whatever between the two disciplines is the risk of concluding that philosophers can generate correct analyses of tiny portions of a subject matter concerning which they may know nothing. This can happen; it does sometimes happen, unfortunately. A strictly factual historical grounding is no crutch for the validity of a piece of reasoning. But neither need such grounding detract, especially when we reflect more intently on the fact that most works in history (as against chronicle) are not strictly factual, and most works in philosophy (as against symbolic logic), are not strictly analytical. The hypotheses at the beginning of the last paragraph are thus unfulfilled conditionals. Given the choice between factually well-based analytical validities and factually shallow analytical validities, what argument could possibly recommend the latter? Great historians of science (e.g., Duhem, Koyré, and Hall), as against mere chroniclers, operate on a level of philosophical sophistication and analytical insight which places them far above the fact-grubbing antiquarians with whom histories of science are sometimes uncritically linked. The premissory platforms on which such distinguished historians build their studies are themselves constituted of analytical planks and philosophical nails which enform and strengthen the corpus of their scholarship.

There is a comparable moral to be drawn concerning "ordinary" philosophers and general History of Philosophy.

V

It is a familiar yet offensive practice of the Genetic Fallacy for an historian of philosophy to slice into some lively discussion of a genuine philosophical perplexity (e.g., other minds, statements about the future, the paradox of analysis, etc.), proceeding then to suffocate that discussion by smothering it under layers of exegesis from the history of thought. To the extent that such references contribute to the prior conversation, the analytical points alone (the conceptual distinctions, the arguments), could have sufficed, without all the learned paraphernalia. To the extent that recitation of the paraphernalia serves some additional end, it is out of place within the restricted philosophical discussion before the group. Saying this is only to reiterate the signal point in announced detections of the Genetic Fallacy. If in 600 B.C. Pythagoras proved that P, then the germane contribution to a strictly philosophical discussion is "It has been proved that P," or just 'P'. The "in 600 B.C." adds no grist whatever to the analytical mill.

On the other hand, some discontent with the "triviality" and "parlor game" features of "*mere* philosophical analysis" has developed in recent decades. One should not question that certain linguistic improprieties can signal serious philosophical errors. Indeed, in this year of grace that is an unquestionable truth. But there is a qualitative difference between being reprimanded for such an impropriety by one who is altogether lacking in knowledge of the History of Philosophy, as against the same riposte being made by a philosophical scholar. For all the rigor and precision of the former's onslaught, it cannot but be felt that the mere word-watcher may miss the fuller dimensions of the problem at hand. He seems often not to grasp the real perplexity; his objective is simply to skim over the surface of the problem, cleaning up minor conceptual tangles—without even perceiving the connections between *this* discussion and thousands more like it, before and after the present time. Being well-informed in the History of Philosophy means having at one's command the facts of the history of thought, the *de facto* details of other people's thought. These facts provide depth and a comprehension, a conceptual framework, and a significance to the analytical endeavour. In the absence of such a background one may seem merely to be playing at philosophy. This, even though the strictly factual components within the historical allusions must remain inferentially independent of any assessment of the validity of a philosophical argument. In practice, however, one's knowledge of the history of philosophy is never just a mental compendium of factual propositions. It usually consists in "mixed" commitments such as: "In . . . B.C. Zeno propounded a paradox concerning motion (i.e., a *genuine* paradox)"; "In . . . A.D. Augustine articulated a (genuine) philosophical perplexity concerning the nature of evil"; "In . . . A.D.

Anselm set out the Ontological argument"; "In ... A.D., Russell reduced mathematics to logic"; ... etc. Such pronouncements are a far cry from any strictly factual utterances concerning events in 1066, 1543, 1687, 1776, 1859, and 1945.

A scrutiny of the Genetic Fallacy, then, can indirectly harbor a message for us all. If its moral is taken as coarse and untempered we may refuse altogether to interrelate distinct levels of discourse. We will hold statements of fact so aloof from the conclusions of analysis that never the twain can meet. This division would be unrepresentative of the best in either history or philosophy, for the former would be "chronicle merely" while the latter would be "analysis merely."

Without grossly committing the Fallacy, and without crudely avoiding it, we might strive to enrich levels of discourse by "mixing" them judiciously, in the interests of knowledge and truth, significance and illumination. For History of Philosophy without (doing) philosophy may be factually full, but it is conceptually dark; philosophy without History of Philosophy may be transparently clear, but factually insignificant and ungrounded—irrelevant to most problems most people actually have, whether they be students or scientists, scholars or statesmen. We must seek to avoid the blindness and the emptiness, without offending reason thereby.

Our expositional problem remains, then: How to "blend" facts and analysis, how to "mix" them, without committing any form of the Genetic Fallacy. This, I concede, is a problem for me now just as when I began this paper years ago. For statements of fact and the statements within analyses do inhabit different levels of discourse; they are inferentially independent. Seeking to link them essentially and directly *via* argument is to flirt with the Fallacy. But to cleave them sharply apart methodologically beckons blindness or emptiness. What kind of a "mixing" or a "blending" is it which can allow us to avoid the blindness and the emptiness, and yet also avoid The Fallacy? What kind of a "mixing" is it which can make The History of Philosophy relevant to *doing* philosophy? The answer must lie in the philosophically-pregnant disclosure which the best historians of thought employ with ever-increasing precision and sophistication. For it is not a merely calendrical matter to have established the date on which Pythagoras' theorem was *proved*. It is not a simple matter of fact that a *Copernican Revolution* was initiated in 18th century philosophy. We are not obliged to assent to the claim that Russell and Whitehead *reduced* mathematics to logic, as we may be to the claim that Lavoisier reduced water to hydrogen and oxygen. Only the philosophically-trained can comprehend such claims, and that very training renders such men the least likely to accept the claims as fact, pure and simple.

Yale University

Received April 6, 1966

American Philosophical Quarterly
Volume 4, Number 2, April 1967

III. LOGIC AND THE PROBLEM OF EVIL

DEWEY J. HOITENGA, JR.

WITHIN recent years the ancient charge has been renewed against theism that, because its central propositions are inconsistent, it is, taken as a whole, false. In one remarkable instance, indeed, the claim was made that the central proposition itself, "God exists," is self-contradictory, and therefore can not be true.[1] The charge with which the present essay is concerned, however, involves more than one proposition and is referred to as the problem of evil. J. L. Mackie states the charge confidently and succinctly: "He [the theologian] must now be prepared to believe, not merely what cannot be proved, but what can be disproved from other beliefs that he also holds."[2] Defenders of theism have rejected the charge with equal confidence. Nelson Pike, for example, writes: "This problem, of course, does not exist. Since there is no logical contradiction between the theological and the ethical theses, there is nothing here to be recognized as rationally offensive."[3]

That philosophers will exhibit such confidence on opposite sides of an issue ostensibly involving a mere question of logical consistency is surprising. It suggests, at the very least, that the real issues dividing them are different. Part of what I wish to show in the following account is that this is so—that the problem of evil is not (except superficially) a question of the formal consistency of propositions, but a question of the meaning of the terms those propositions involve. The problem of evil challenges, I shall argue, our fundamental conceptions of good and evil, freedom and causation, the divine and the human; not our ability to detect or construct sets of propositions that are formally consistent or formally contradictory. In the course of this main argument I will have occasion to point up the logical dissimilarity between the traditional "higher good defense" and the "free will defense" of some recent theists—a dissimilarity which I add to a case I wish to make against the latter defense. I shall return, in the conclusion, to a brief statement of what I regard as the more general significance of the main argument concerning the *logic* of the problem of evil.

The disagreement between sceptics and theists should first be set forth in some detail. Sceptics hold that the existence of a God who is omnipotent, omniscient, and all-good or omnibenevolent logically rules out the existence of evil; and defenders of theism deny this. The most familiar formulation of the sceptic's case is repeated in the following lines by Professor Aiken:

> If there is a being that is at once almighty, omniscient, and perfectly good, it will be both able and willing to prevent evil. On the other hand, if something is evil, it must be concluded that there is no such being, since a perfectly good person would prevent it if he could, and an almighty and omniscient being could prevent it if it would. Either, then, there is no such being or nothing is evil.[4]

The formal validity of the argument requires the additional premiss that if a being is both able and willing to prevent evil, no evil will exist. Though this premiss is taken for granted in Aiken's statement, it is not the premiss, I suppose, which defenders of theism will care to dispute. They will instead be seen to deny either that an almighty, omniscient, and omnibenevolent being is *able*, or that he is *willing*, to prevent evil; or both.

If, however, God *could not* prevent evil, he must be powerless in the presence of the possibility of evil; while if he *would not* prevent it, he must have some purpose—a good one, if he is all-good—in allowing it. Thus, given the central beliefs of traditional theism, we may distinguish two "problems of evil": first, an etiological problem: Whence evil? That is, how did it come to exist? What conditions made it possible in the first place?; and secondly, a teleological problem: Why is there evil at all? That is, given evil, and the nature of God as this is understood in theism, what purpose does it serve, in consideration of which its occurrence may be justified? I shall discuss the latter problem first, for I hope to bring the theist position on the justification problem to bear on some recent theist approaches to the problem of the origin of evil.

---

[1] J. N. Findlay, "Can God's Existence Be Disproved?" in A. Flew and A. Macintyre (eds.), *New Essays in Philosophical Theology* (New York, 1955), ch. 4.

[2] "Evil and Omnipotence," *Mind*, vol. 64 (1955), p. 200.

[3] Nelson Pike, "God and Evil: A Reconsideration," *Ethics*, vol. 68 (1958), p. 123.

[4] Henry D. Aiken, *Reason and Conduct* (New York, 1962), p. 175.

I

The moral justification of evil invariably turns, in theist literature, on the concept of a higher good, and it may aptly be called, therefore, the "higher good defense." I should like to set forth the necessary ingredients of the "higher good" concept which such a defense assumes. In the article cited above Nelson Pike uses just such a defense to reject the sceptic's charge of inconsistency. His central claim in this article is that "the logic of the term 'good' (if such there be) does not demand that a perfectly good person would prevent evil, even if he could."[5] He argues in support of this claim as follows:

> A perfectly good person could allow evil, or for that matter be himself an evildoer—though it be avoidable, providing only that there be some other "good reason" for his tolerance of evil or his evil-doing. The point to be emphasized here is strictly semantical. The term "good" can apply, within the limitations of its proper usage, to an evildoer so long as it is possible to cite a motive or other factual condition which renders *moral blame* for the evil action inappropriate. Such a motive or condition could, of course, eliminate the property of moral *blame* without at the same time alleviating moral *responsibility* for the act in question. The father is morally responsible for the pain of the punished child, yet we would not feel it appropriate to blame him for his action if it were done with the child's ultimate good in mind.[6]

An analysis of this argument and of the example it employs will exhibit the ingredients of the notion of a higher good as it is required for the theist's defense. These ingredients are four:

1. Such a good is dependent upon the specific evil which it justifies. So in the example, it is assumed that the punishment of the child will constitute a condition for his reformation. All such higher goods, that is, are causally related to the evils which they justify, and knowledge of such causal relationships is involved in the responsibility and blamelessness of the agents who inflict or tolerate these evils.

2. Such a good is not only dependent upon the specific evil, but it is obtainable in no other way which excludes the evil. So in the example, it is assumed that punishing the child constitutes a necessary condition for bringing about his ultimate good. All such goods, therefore, are causally related to the evils they are used to justify as their necessary conditions, and knowledge of these conditions as the ones necessary for the goods in question is involved in the responsibility and blamelessness of the moral agents who inflict or tolerate the evils.[7] If this were not so, some lesser evil or positive good should be sought as the means to the good end.

3. Such a good outweighs in value the disvalue of the evil on which it depends. So in the example, it is assumed that the reformation of the child is more of a value than the pain he must suffer in punishment is a disvalue. All such goods, that is, are better than the evils needed to bring them about are worse—they are "worth the price that has to be paid." Otherwise a father could be blameless for inflicting extreme torture on his child to prevent him from repeating a minor misdemeanor. Thus knowing the values of the goods achieved in comparison with the disvalues of the evils required is another part of the responsibility of those who inflict or tolerate such evils.

4. Finally such a good must outweigh in goodness any alternative good that could be achieved by different, non-evil, means. So in the example it is assumed that the type of reformation which only punishment can bring about is a greater good than any alternative type of reformation implemented by some other means—say permissiveness. Otherwise the father who had knowledge of and capability for each set of means and end would hardly be blameless for choosing reformation by punishment over reformation by permissiveness. All such goods, therefore, must not only outweigh in goodness the evil of the evils on which they depend; they must also outweigh in goodness the goodness of any other goods which could be attained without the evils in question.

This last ingredient is so important in the sceptic's case that I should like to illustrate it with another example—the crucial one of suffering pain in situations where it is not inflicted as a justifiable punishment. It has been argued by theists[8] that such pain is nevertheless justified by the higher goods it makes possible, such as virtuous responses like patience and fortitude. The argument is sound, however, only if such virtues are higher goods than alternate virtues, such as contentment and industriousness, which are made possible by a state of health lacking the pain. And if the pain is borne bitterly instead of nobly, then the evil of the pain, far from being justified by a good that outweighs it, is actually rendered more intensely unjustified by the further evil of bitterness that not only depends on it but is invited by it. Mackie considers such instances as constituting a "fatal objection" to the theist defense.[9] They are fatal because there is no good present (at least obviously so) in these instances; yet this is precisely what the justification of any evil, on the theist's terms, requires.

---

[5] Pike, *op. cit.*, p. 119.

[6] *Loc. cit.*

[7] Pike elsewhere provides some examples to show how the good may precede the evil it justifies as well as follow it. (*God and Evil* [Englewood-Cliffs, 1964], p. 91.) In either case the important feature is that the evil constitutes an unavoidable condition for the occurrence of the good.

There are further difficulties, however, that arise when we actually apply the concept of a higher good, as analyzed above, to the activity of God. For the plausibility of Pike's example—and of all such examples, I would venture—rests upon its appropriateness in a *human* situation. It is of the essence of such situations that (1) evil is already *given* as an ultimate condition which has to be reckoned with in the attainment of value; and (2) blamelessness in a moral agent who inflicts or tolerates an evil is a function of (a) his intention to achieve a higher good over a lower one, (b) his capability to achieve such a higher good, and (c) his knowledge both of the factual (causal) relationships and of the comparative values and disvalues of the goods and evils involved. Now such examples always show how human moral agents can be spared from blame in a situation where evil already exists and has to be taken into account. For only such a situation makes possible having sometimes to choose goods dependent on those evils. But Pike, and according to him all theists, deny that evil exists in any universe as an ultimate condition. If they are correct in this, as I believe they are, then the conditions under which God comes to create, sustain, and rule a universe are different from the conditions of human action in a most essential way. But if the examples get their force from a situation which is inapplicable to God, they will hardly support the theist's case.[10]

If the existence of evil is not an unconditioned necessity, nor an arbitrary accident (theists would reject this, it seems to me, with equal vigor), it must be a conditioned necessity. It must be conditioned by what is ultimate, namely, by the highest good; and this is the goodness of God. What this means is that the theist, in providing for the justification of evil, provides also for its necessity. For the justification of the evil that exists is made in terms both of an assertion that some good depends upon it, and of a denial that any higher good not depending upon it is available instead. From this it follows that evil is not only justified but necessary. For if evil *may be* justified, then, if God exists and evil exists (both of which the theist grants), it *is* justified; for if it were not, the existence of God and the existence of evil together would still constitute the contradiction which the conditions of justified evil were put forth to remove. But if evil is justified, then it is also necessary, for it is the necessary condition for the good which, since God is all good, is therefore willed by Him. Given then His goodness, together with his omniscience and omnipotence, evil is inevitable.[11]

In view of all this the *meaning* of the term "good," that is, its logic, its entailment rules, will be predetermined by the admission of the fact of evil. And this Pike does admit:

> The important point has already been made: viz., the assumptions or premises of our whole discussion guarantee (logically) that there is *some* good reason for evil in the world. The theologian could condemn as erroneous all of the theodicies just considered and could even reject all extant theodicies as theoretically or existentially inadequate. So long as he insists there is *some* good reason for evil, one which remains only to be discovered by future investigation, he can accept both the theological and ethical theses without threat of contradiction. . . .
>
> Within theology, the theological thesis is accepted prior to and is indeed *axiomatic to*, the rational consideration of the status of evil in the natural world.[12]

---

[8] For an example see C. S. Lewis, *The Problem of Pain* (New York, 1947), chs. 6 and 7.

[9] *Op. cit.*, p. 207.

[10] As a further objection against Pike's example of the father punishing the child the sceptic might insist on the more accurate analogy of the father providing a moral temptation for his child which he knew the child would not resist.

[11] As Leibniz argued: *Theodicy*, sec. 25.

[12] Pike, *op. cit.*, pp. 121 ff.

I do not see how Pike avoids exposing himself here to the very charge he levels against both theologians and philosophers in regard to the term "evil": that of defining a term, in the present instance, "good," according to one's own interests, thereby solving nothing.[13] But that is only one difficulty. As Pike understands it, the only problem for a "practicing theologian"—one who accepts "both the theological and ethical theses as *premisses* or *axioms*"—is not to resolve a contradiction but to *discover* something, very much like a scientist who tries to discover something that will render an hypothesis which a particular experiment that was designed to confirm fails to do. But the sceptical question now forces itself: does he know, from thinking about goodness as he confronts it in his human situation, what he is looking for in God's?

It should be clear by this time that the issue between sceptics and theists occasioned by the existence of evil and the belief in God turns not on the consistency or inconsistency of a set of propositions, but on the meaning of "good" and on what "all-good" in the proposition attributing this quality to God entails. For the sceptic a state of affairs can be good, or a higher good, only if it is without evil. Introduce whatever so small amount of evil possible into this state of affairs, and it becomes less than good, less good than a state of affairs containing no evil whatsoever. It may occasion different sorts of goods that depend on the evil introduced, but these goods are not so great as those that could be achieved without any evil at all. Hence the sceptic's entailment rule for the term "good":

> These additional principles [connecting "good," "evil," and "omnipotent"] are that good is opposed to evil, in such a way that a good thing always eliminates evil as far as it can, and that there are no limits to what an omnipotent thing can do.[14]

For the theist, however, the introduction of evil into a good state of affairs need not lessen the goodness of the whole; indeed, he holds that such an introduction can occur leaving the whole better than it was. Hence his entailment rule for the term "good":

> The logic of the term "good" . . . does not demand that a perfectly good person would prevent evil in the world, even if he could.[15]

These are, I submit, two simply irreconcilable views on the nature of goodness.

Let us name the first view, the one which sustains the sceptic, the "Incompatibility View," for according to it the opposition between good and evil is simple and absolute. The one always, by its very nature, tends to prevent or destroy the other. According to such a view, a good world can only be one that contains no evil. Any world in which evil exists is thereby not the good world it could be if there were no evil in it at all; and it certainly could never be *better* than such a simply good world. As for those good things that depend on evil, they indeed represent the best that can be had in a situation already partly evil, but in so far as they do not prevent or eliminate the evil on which they depend they can not really be called good at all, for good by definition is contrary to evil.

This view of the nature of good and evil is drawn, I believe, from our universal moral experience; more specifically from the content of moral obligation: the unconditioned imperative to avoid and remove all evil. Bearing pain, for example, is never a good state of affairs, for we are always morally obligated to prevent or remove it in so far as we are able. But then it can never be (or contribute to) a good state of affairs, for we are never obligated to frustrate a good state of affairs. The fact that we are sometimes unable to remove such an evil state of affairs without creating a greater evil in its place serves only to show the limitations of our abilities and/or knowledge and thereby to spare us from moral blame. It does not reduce the evil of the pain-bearing. That a case of pain-bearing might become a predominately good state of affairs when done nobly will not count against the argument—unless we have an obligation to make possible such occasions of noble pain-bearing by inflicting or tolerating the pain required. But this no conscience will approve.[16]

In other words, if we are unable to remove any evil through ignorance, lack of ability, or because we must in removing it create greater evil, we are excused from the obligation. Such excuse, however, is due to our limitations with respect to fulfilling the obligation, and not to any weakness or condition in the obligation itself. Our blamelessness is owing not, that is, to the evil's being any less evil and therefore no longer to be opposed. Indeed, if our ignorance were overcome, our ability increased, and the threat of inducing greater evil removed, we would no longer be excused from the obligation to remove the evil. Now God is by definition a being who *is* unlimited in knowledge and power. Thus if he is also a moral being whose goodness is understood on an analogy with the goodness of human beings who fulfill their obligations, and if he exists, his existence must rule out the existence of evil. But evil exists, and hence the contradiction in theism which constitutes the case of the sceptic. To understand God and evil in terms of our own moral struggle is to raise some doubt in regard to the existence of God. The sceptic prefers this doubt to a doubt of his own conscience.[17]

---

[13] *Ibid.*, pp. 118 ff. For two classic illustrations of the point see Augustine, *Enchiridion*, sect. XCVI, and Leibniz, *op. cit.*, sec. 10.

[14] Mackie, *op. cit.*, p. 201. In Aiken's words: "But any perfectly good person, so far as he can, will do good and prevent evil" (*op. cit.*, p. 175).

[15] Pike, *op. cit.*, p. 119.

[16] In A. W. Price's words, speaking of the healings in St. Stephen's Episcopal Church in Philadelphia a few years ago: "It's true that pain can be used for victory through the cross, but per se, it's evil. . . . Jesus never told the suffering who came to him that 'I won't heal you because suffering is good for your soul.' " Reported in *Christianity Today*, vol. 8 (1963), p. 39.

Let us name the second view, the one that sustains the theist, the "Compatibility View," for according to it the opposition between good and evil is not simple and absolute. Good, instead of always tending by its nature to prevent or destroy evil, sometimes includes and even depends upon it; evil, instead of always tending by its nature to undermine good sometimes provides necessary conditions for it. Any world in which evil does not exist may thereby still be a good world; indeed, it may be better than a world without evil. For what makes a world good in this view is not simply how much evil is eliminated by good, but how much good is made possible by the evil it contains, in addition to the good that exists independently of this evil.

This view is not drawn from the nature of moral obligation but from the Judaeo-Christian religion which testifies to a God whose will is not only the ultimate source of all being but also the final criterion of what is good and just. But it is a will that not only brings about evil but even issues commands to human beings that on occasion are in conflict with those of conscience.[18] Since it is the will of God that is the final standard, not the judgments of men, and since men are always to submit their will and judgment to the revelation of this divine will, any contradiction between the divine will and the human is, in principle, removed. Thus the defense of theism rests on an identification of goodness with the will of God—an identification which, given the existence of evil, raises some doubt in regard to the deliverances of the human moral sense. The theist prefers this doubt to a doubt of the existence of God.

II

The "higher good defense" asserts that God would not have prevented evil—even if he could have—because if he had, he would have prevented the achievement of a higher good. But if God would not have prevented evil, it must have been a distinct possibility; for if it had not been, the occasion of his willing it for the sake of higher goods could not have arisen. Now the possibility of evil, and for that matter its actual origination, has been attributed by Christian theologians since St. Augustine to the freedom of the created will. The type of theistic defense I shall now examine has arisen around the concept of this freedom and has appropriately been called the "free will defense."

It is important to notice at the outset that a defense of theism oriented to the freedom of the will might turn out to be nothing other than a form of the higher good defense. It might be argued, for example, that the creation of a world of freedom with its attendant possibility of moral evil is a higher good than the creation of a world without such freedom. Such an argument would not be a *specifically* free will defense, however; it would be indistinguishable in its logic from the teleological justification of evil. It would assert that there is a good (freedom of the will) that entails the possibility of evil, but which outweighs in goodness both whatever actual evils might arise from that possibility and any alternative creation which eliminated that possibility.

The "free will defense" as it has been developed by some recent writers does not take this form. It argues rather that God *could* not have prevented evil in the creation of free wills even if he *would* have. That such a defense raises a quite different issue from that in the higher good defense may be seen from the following: Suppose, contrary to the higher good defense, that there were no higher good in terms of which evil could be justified. Evil might still arise through God's *inability* to prevent it. And God would be unable to prevent it if it is free wills he wants to create—and it happens that he does want to create such wills. The crucial proposition here is that concerning God's inability to create a world of freedom and guarantee its goodness at the same time. In any full-blown theodicy, of course, it must be held that such a world is a better one, in order to preserve God's goodness. For if God *could* not prevent evil in creating free creatures, then there is no need to hold that he *would* not have prevented it even though he is all-good, if and only if the creation of free creatures is a greater good than all alternative creations or no creation at all. In this way any free will defense requires a higher good defense, and most discussions of freedom in the context of the problem of evil take the value question into view.[19] But I wish to distinguish, as I think the advocates of the free will defense have distinguished, the problem of how such freedom is to be *conceived* in a theistic context from the problem of how freedom as conceived *contributes to* this world's being better than some other kind or none at all. Mackie claims that the way in which freedom is conceived involves the solution of evil in terms of a higher good.[20] I think he is right about that; but it also involves, as I hope to show, the conception of divine omnipotence. I shall hold that God has power to do what he wills (and what he wills is good), even if such power is conceptually peculiar in so far as it includes the determination of created freedom—just as his goodness is conceptually peculiar also in including evil. The advocates of the free will defense hold, on the contrary, that God wills to do what He can (and this is good); but such power is not made conceptually peculiar by including a determination of created freedom—no more than goodness, according to the sceptical view, is to be made conceptually peculiar by having to include evil.

---

[17] For an alternative position which sacrifices divine goodness to maintain both divine existence and human conscience, see A. E. Taylor, *Elements of Metaphysics* (London, 1961), pp. 396 ff.

[18] *Genesis* 22; *Joshua* 8.

The central logical claim in the free will defense, then, is that the concept of a free person being guaranteed infallibly good by God is self-contradictory. It is well-stated by John Hick as follows:

> To be a person is to be a finite center of freedom, a (relatively) free and self-directing agent responsible for one's own decisions. This involves being free to act wrongly as well as to act rightly. The idea of a person who can be infallibly guaranteed always to act rightly is self-contradictory. There can be no guarantee in advance that a genuinely free moral agent will never choose amiss. Consequently, the possibility of wrong-doing or sin is logically inseparable from the creation of finite persons, and to say that God should not have created beings who might sin amounts to saying that he should not have created people.[21]

The arguments put forth in support of the inconsistency of saying that an action is both freely performed and guaranteed good usually presuppose the identification of "free" with "morally indeterminate." I shall oppose the position taken in the free will defense by arguing first that consideration of two important "theistic paradigm cases" must force the free will defender to give up his claim; secondly, that his claim involves him in serious difficulties at other points in his belief; and finally, that he already has a model in theistic belief for the conceptual peculiarity of "free and determined" which he wishes to avoid.

There are two theistic paradigm cases in which a person is regarded as wholly free yet entirely secure in goodness. First of all, there is the case of God himself. Theists certainly agree that God is a perfectly good, yet also perfectly free being. For the perfection or infallibility of His goodness is part of what is intended in calling Him "all-good," while perfect freedom is but another term for his omnipotence. Yet no theist fears that in making the assertion of God's omnipotence and all-goodness he is in danger of self-contradiction.[22] The fact that part of theistic belief contains precisely a model for a perfectly good and perfectly free agent is in striking contrast to the insistence of free will defenders that these are strictly incompatible.

---

[19] Thus Alvin Plantinga in his summary of the free will defense does not fail to mention the following proposition: "A world containing creatures who freely perform both good and evil actions—and do more good than evil—is more valuable than a world containing quasi-automata who always do what is right because they are unable to do otherwise." "The Free-Will Defense," in Max Black (ed.), *Philosophy in America* (Englewood-Cliffs, 1965), p. 205.

[20] Mackie, *op. cit.*, p. 208.

[21] *Philosophy of Religion* (Englewood-Cliffs, 1963), pp. 41 ff. Plantinga's account is essentially the same: "Now God can create free creatures, but He cannot causally or otherwise determine them to do only what is right; for if He does so then they do not do what is right *freely*" (*loc. cit.*).

[22] "The freedom of the Almighty hinders not his being determined by what is best" (Locke, *Essay Concerning Human Understanding*, Bk. II, ch. xxi, sec. 50).

The free will defenders might reply that their thesis of inconsistency is intended only as an account of human freedom, that the case with God is understandably different. Since man is a creature he is neither all-good nor entirely free; in fact he is limited in many ways that God is not. But I do not see why the difference between creator and creature might not prescribe different types of moral perfection—that appropriate to God and that appropriate to man—without thereby entailing the specific characteristic of the corruptibility of the latter. What has not been shown is that a moral determination of human freedom by human goodness analogous to the moral determination of divine freedom by divine goodness is impossible.

I doubt whether the difference between God and man can be so stated as to rule out the consistency of "free" and "guaranteed good" as predicated of man. For this is just the description of man's destiny in a future life that theists provide in a more elaborate statement of their belief. Man's moral perfection in that life is presented as quite secure, yet with no loss to his essential humanity or freedom. This description provides the second paradigm case. According to it freedom and infallible moral perfection can (and will) be, as Augustine says, a "gift of God."[23] But if God cannot do the logically impossible, as all parties to the dispute insist, then human freedom and infallible moral perfection are quite consistent characteristics. But then the reason for their not having been both given to man in his creation can not be their logical incompatibility as the free will defenders claim.

Let us look at some typical arguments which these defenders of theism use, nevertheless, to establish such incompatibility. One argument, not conspicuous in the recent discussion, might be an inference from man's present moral condition to that in which he was originally created. If evil choice is now actual, the argument might go, then it must have been possible in the created state; and if it was possible in the created state, this state could not have been one in which goodness was guaranteed. But there is no obvious deduction from the fact that the Fall did occur and was therefore a possibility built into human freedom to the claim that it *had* to be built in. This is the simplest reply to the argument, but more can be said. Even if freedom as it *was* created was not infallibly good, it was nevertheless, on the theistic account, good enough to make evil moral choice deeply puzzling.[24] Efforts have been made in the interests of the free will defense, to circumscribe the goodness of the original freedom of man. C. S. Lewis, for example, distinguishes between the self (which is capable of an idolatrous choice) and its dispositions (which in the created state were good).[25] But what is this self abstracted from its moral disposition and practice? Lewis answers: It is "the 'weak spot' in the very nature of creation, the risk which God apparently thinks worth taking." This answer (and similar ones that have been offered)[26] suggests, however, that the distinctions emerge for the sake of explaining the possibility of moral evil, rather than from an analysis of the nature of man as he was created—good and upright. Certainly the goodness of the self's disposition provides a basis for the moral judgment that the fall is, in Lewis' word, "heinous"; but it also provides a "strong spot" in terms of which that fall must appear logically very puzzling. For if the human will as created by God was indeed, as Lewis claims, "wholly disposed, though not compelled, to turn to God," how can the will so disposed ever be said to *fail* to turn to God, without suggesting contradiction?

Another argument, one that is conspicuous in the recent discussion, is that any guarantee of the moral quality of human actions on the part of divine agency undermines the freedom in which those actions are performed. This freedom is never, at least *prima facie*, independent of God, for the doctrine of creation asserts that God is the ultimate origin of all things that exist besides himself. But if freedom originates in a creative act of God, and the occurrence of evil is traced to freedom, we clearly must trace the occurrence of evil ultimately to God. Free will defenders reply by distinguishing between two senses of freedom: freedom as it is created by God and freedom as it is the source of human acts. Thereby they can also undercut any argument tracing evil to God, if that should happen to be part of their motivation. Alvin Plantinga, for example, argues:

---

[23] *City of God*, Bk. XXII, ch. 30.

[24] "In the developed doctrine, then, it is claimed that Man, as God made him, was completely good and completely happy, but that he disobeyed and became what we now see. . . . We have no idea in what particular act, or series of acts, the self-contradictory, impossible wish found expression" (Lewis, *op. cit.*, pp. 60, 68). Augustine's perplexity is well known; see *City of God*, Bk. XII, ch. 6.

[25] Lewis, *op. cit.*, pp. 68 ff.

[26] St. Augustine refers on one occasion to the created will as "some intermediate quality which can be called neither folly nor wisdom" (*On Free Will*, Bk. III, ch. xxiv, sec. 71; see also Bk. II, ch. xviii, secs. 47–50). And Leibniz declares: "For we must consider that there is an *original imperfection in the creature* before sin, because the creature is limited in its essence" (*op. cit.*, sec. 20).

> And (5b) ["God creates free men and brings it about that they always freely do what is right."], of course, is not consistent; for if God *brings it about* that the men He creates always do what is right, then they do not do what is right freely.[27]

The only way this argument can hold is through the distinction of two senses of freedom just mentioned: one according to which God creates free men (where freedom is clearly dependent upon God); and the other according to which freedom issues in morally right acts (where freedom's moral determination by God is rejected: "Whether the free men created by God would always do what is right would presumably be up to them; for all we know they might exercise their freedom to do what is wrong."[28])

Three objections can be made against this line of argument. First of all, traditional Christian apologists, as far as I have been able to determine, do not find anything logically inconsistent in the concept of a free and infallibly good human being. In his dialogue *On Free Will* Augustine has Evodius ask: "But don't you think . . . that, if it [free will] was given for the purpose of welldoing [which Augustine has asserted], it ought not to have been possible to convert it to sinful uses?"[29] It is interesting to notice that Augustine does not reply, as the modern free will defenders do, that it would have been logically impossible for God to have prevented this conversion to sinful uses. Since this would have been the simplest answer to give, the fact that he does not give it suggests that he did not believe it. He replies instead that the possibility of this conversion to sinful uses must have been rightly given to man since God is its giver, and God is good: "If we allow that God gave it, ought we to say that he ought not to have given it? . . . If it is certain that he has given it, we ought to confess that, however it was given, it was rightly given."[30] And in another passage in *City of God* he writes:

> For who will dare to believe or say that it was not in God's power to prevent both angels and men from sinning? But God preferred to leave this in their power, and thus to show both what evil could be wrought by their pride, and what good by His grace.[31]

This is an even clearer rejection of the contemporary free will defense argument from logical impossibility. Though it includes an indeterministic characterization of created freedom, it defends such a characterization by reference to the goodness of divine intentions, not to the logical limitations on his omnipotence.

Secondly, by affirming the independence of the moral character of the created will from God in order to preserve its freedom, the argument proves too much in that it prevents God's authorship of the good that men do. This flies in the face of the universal attribution by theists of all good, including the good men do, to God:

> For when I am evil, then to confess to Thee, is nothing else than to be displeased with myself; but when holy, nothing else than not to ascribe it to myself: because Thou, O Lord, blessest the godly, but first Thou justifieth him when ungodly.[32]

Thirdly, the distinction between freedom as created by God and freedom as it issues in moral acts appears to be made, in Deistic fashion, between the mere existence of freedom and its

[27] *Op. cit.*, p. 210. See also A. Grave: "There is, by definition, a logical impossibility in God's making him freely choose the good on any occasion" ("On Evil and Omnipotence," *Mind*, vol. 65 [1956], p. 260).

[28] Plantinga, *op. cit.*, p. 211. Descartes distinguishes between the "faculty" of freedom and its "operation" in *Meditations*, No. IV.

[29] *On Free Will*, Bk. II, ch. i, sec. 4.

[30] *Loc. cit.* Compare Descartes: "I perceive, nevertheless, that it would have been easy for God to contrive that I would never make mistakes, even though I remained free and with limited knowledge" (*op. cit.*). And Leibniz: "But since he has permitted vice, it must be that that order of the universe which was found preferable to every other plan required it. One must believe that it is not permitted to do otherwise, since it is not possible to do better. It is a hypothetical necessity, a moral necessity, which, far from being contrary to freedom, is the effect of its choice" (*op cit.*, sec. 124).

[31] Bk. XIV, ch. 27.

[32] *Confessions*, Bk. X, ch. 2.

function (reminiscent of Lewis' distinction between the mere self and its dispositions). But if freedom is one thing and its exercise another, the suggestion is that behind (underneath?) the freedom given by God is another, hidden freedom in man, which may or may not exercise the God-given freedom in a certain way. And this exposes the free will defense to a very forceful objection, viz., that it is bound to set a limit to God's control over the course of human events. But any such limit will conflict with the theist's belief in providence—that all things, including human history and destiny, are under that control. In Augustine's words:

> For He cannot truly be called Almighty if He cannot do whatsoever he pleases, or if the power of His almighty will is hindered by the will of any creature whatsoever.[33]

But if it is true that the moral choices of men are simply "up to" them, as Plantinga says, it would seem to follow that at least some things are not up to God. How then will God accomplish his benevolent purposes? On what ground will the pious man come to trust Him?[34] If the free will defense, then, succeeds in removing any alleged contradiction in God's being said to cause the men he creates always to do what is right, it does so only at the price of introducing another contradiction: one between "all things are contingent upon the will of God" and "some things are not contingent upon the will of God." This is no small price to pay for defenders of theism whose predominant motive is precisely to avoid what they take to be a self-contradiction in the concept of an infallibly good, free person.

Efforts have been put forth—unsuccessfully, as I think—to avoid paying the price of an open rejection of divine omnipotence, even though such a rejection appears to be entailed by the free will defense. A. Grave, although he admits that "part of the meaning of omnipotent is 'always able to control everything'," goes on to distinguish between an act (which is left to be determined by the human self) and its consequences (which are "bent" toward fulfilling the divine purpose).[35] But this will not succeed, for it leaves the most important aspect of the events in question, namely their origin in human choice, frankly outside the very control which is said, nevertheless, to be universal in scope. Pike utilizes the distinction between moral responsibility and moral blame, openly admitting, as we have already seen, the possibility that God is an evil-doer.[36] Though Pike does not discuss (in the article cited) the problem of freedom, he could, by attributing responsibility but not blame for the evil acts of men to God, still preserve the omnipotence and universal control which Grave must compromise.

I should now like to raise an objection to the free will defense which is quite different from all those above. It is, simply, that the defense takes for granted the very entailment rules that are in dispute. For quite clearly there is no *formal* contradiction between "God determines that men always act rightly" and "men act rightly freely." How then is the contradiction established for the sake of the free will defense? In the same way, interestingly, that such contradiction is avoided by others who seek to ruin that defense—by an appeal to the *meaning* of the terms. Compare the following accounts:

> . . . to say that a person could have helped doing something is not to say that what he did was in principle unpredictable nor that there were no causes anywhere which determined that he would as a matter of fact act in this way. It is to say that *if* he had chosen to do otherwise he would have been able to do so; that there were alternatives, within the capacities of one of his physical strength, of his I.Q., with his knowledge, and open to a person in his situation.[37]

> When we say that Jones acts freely on a given occasion, what we say entails, I should think, *that either* his action on that occasion is *not* causally determined or he has performed some other undetermined action which is a causal ancestor of the one in question.[38]

Here are set forth two quite conflicting views of the logic (meaning) of the terms "free" and "determined." In one view (Flew's) "free" is consistent with "determined"; it is held that a given act can be accounted for with both terms without contradiction. In the other view (Plantinga's) "free" is inconsistent with "determined," and thus equivalent to "undetermined"; that is, it is held that a given act can never be accounted for with both terms without contradiction. The heart of the free will issue turns, therefore, on semantic or conceptual, not propositional analysis. Having opened this possibility, Plantinga proceeds to avoid it as irrelevant:

---

[33] *Enchiridion*, sec. XCVI.

[34] "We know that in everything God works for good with those who love him, who are called according to his purpose" (St. Paul, *Letter to the Romans*, 8:28 [R.S.V.]).

[35] Grave, *op. cit.*, p. 261.

[36] Pike, *op. cit.*, p. 119.

[37] "Divine Omnipotence and Human Freedom" in A. Flew and A. Macintyre, *op. cit.*, p. 150. For a similar view, see C Ducasse: *A Philosophical Scrutiny of Religion* (New York, 1953), p. 378.

[38] Plantinga, *op. cit.*, p. 207.

> The Free Will Defender claims that the sentence "not all free actions are causally determined" expresses a necessary truth; Flew denies this claim. This strongly suggests that Flew and the Free Will Defender are not using the words "free" and "freedom" in the same way, and I *think* it is Flew who is using these words in a non-standard, unordinary way. But we need not try to resolve that issue; for the Free Will Defender can simply make Flew a present of the word "freedom" and state his case using other locutions.[39]

This will not do; at least some concrete example ought to be provided for "using these words in a standard, ordinary way"—more accurately, for using these words in a standard, ordinary *theist* way. This is what Flew intended to do in bringing forth his "paradigm case," analysis of which, according to him, reveals the compatibility of the usage of "free" and "determined" even outside a theistic context. This was also my point above in advancing the paradigm cases which, by contrast with Flew's, are taken from within theistic belief. To reply that the usage of the terms in the instances cited (God, and man in his future life) is unordinary would only be tendentious; it certainly is ordinary for many traditional theists. Instead of taking for granted one logic of the terms "free" and "determined" and building their case upon it, the free will defenders should first defend the logic itself, from within theism as a whole.[40]

There appear to be, then, two simply irreconcilable views on the free will issue, very much as we discovered to be the case with the nature of good and evil. Again we can name one of the views, the one which sustains the free will defense, the "Incompatibility View," since according to it freedom and determinism are contrary concepts. If an act (right or wrong) is performed freely, it is thereby performed with no prior divine guarantees as to its rightness or wrongness; for there to have been such a guarantee would have ruled out its being a free act. This view is derived, I would venture, as a generalization from those experiences in which we find a force outside our control preventing us from effecting something by our own will. In such instances our "being determined" or "caused" entails our "being unfree." In those cases, however, where such a conflict between being determined and being free is not *noticed* or *felt*, we easily suppose that the action in question is within our control and *thereby* not determined. Indeed, when we take a broader view of ourselves as part of a larger whole this supposition is even apt to raise a doubt in our minds in regard to our sense of the universality of the causal order of things; yet from the moral point of view we prefer that doubt to a doubt of our moral freedom and responsibility. Nevertheless, from the fact that the presence of something is not noticed or felt there is no valid deduction to its absence, and hence the generalization—though perfectly understandable—is really groundless.

The other view, the one that sustains such theists as Augustine, Calvin, and Leibniz and such non-theists as Flew, may be called the "Compatibility View," since according to it "being free" and "being determined" are not necessarily contrary concepts at all. To regard an act (right or wrong) as performed freely is not thereby to rule out its being previously guaranteed (by God) as to its rightness or wrongness. This view is derived, as it seems to me, from a concern to take full account of all the beliefs that bear on the analysis of an act. In a non-theological context these beliefs are especially two, both of which undergird Flew's defense of the compatibility view: freedom of the will, and the universality of the causal order. In a theological context they are also two: freedom of the will and divine providence. Yet both Kant (in a modern scientific context) and Augustine (in a theological one) regarded "being free" and "being determined" as compatible accounts of the same human act.[41] Such a theological position is apt to raise a doubt in regard to our sense of being masters of our fate, but theists have been known to prefer that doubt to doubting the sovereignty of God.

---

[39] *Loc. cit.*

[40] Some theists, indeed, even speak of *evil* choices being free yet determined: "We have also laid down a distinction between coaction and necessity, from which it appears that while he sins necessarily, he nevertheless sins voluntarily" (John Calvin, *Institutes of the Christian Religion*, Bk. II, ch. iv, sec. 1).

[41] See Kant *Foundations of the Metaphysics of Morals*, Part III; and Augustine, *City of God*, Bk. V, ch. 9.

I can now restate my last objection to the free will defense in a more positive way, as well as bring to bear on it the logic of theism as it is exhibited in the "higher good defense." Our examination of that defense revealed that theists mean by "good" something that can include evil, while sceptics mean by it something that cannot include evil. We also saw that the theistic meaning of "good" is applicable to human action only in virtue of human limitation in respect to knowledge and power, but that its applicability to divine action—since that action occurs with no limitation of knowledge or power—derives instead from the admission that evil as a matter of fact exists. This, we concluded, resulted in a mysterious sense of goodness as it is used to describe the being and doings of God.

An examination of the free will defense, on the other hand, revealed that its advocates mean by "cause" or "determine" something that cannot include "allowing to act freely," as opposed to many traditional theists who hold that "cause" or "determine," at least when used of God, can include "allowing to act freely." We also saw that the former view accords with an understandable (if groundless) interpretation of freedom, while the latter, though required by theistic belief in providence, yields a mysterious sense of causality altogether compatible with freedom. My objection, then, to the free will defense, is this: If theists generally require—or at least do not shrink from—for the sake of defending divine goodness in the face of evil, a mysterious kind of "higher goodness" that includes evil, what hinders some of them (by contrast with others) from developing a concept of divine power in the presence of human freedom, the essence of which would be a similarly mysterious kind of "higher causality" that includes such freedom? That is, if theists generally are "Compatibilists" on good and evil, why do some of them hold out for an "Incompatibilism" on freedom and determinism? For from a human point of view it would seem that the compatibility of the latter is no more absurd than that of the former. Is the mystery of a sovereign God who allows some of his creatures freedom of choice without compromising his omnipotence more baffling than the mystery of an all-good God who allows evil because of this goodness?

* * *

The review in this paper of two possible views (a compatibility view and an incompatibility view) on two different issues involved in the problem of evil (the nature of good and evil; and the nature of being free and being determined) reveals that there are four possible approaches to the problem of evil as a whole. First there is a consistent "incompatibility" approach which would hold to the incompatibility view both with respect to good and evil and with respect to freedom and determinism. I think Mackie comes close to this approach in his account, since he holds that good by its very nature tends to eliminate evil, and that genuine human freedom is possible only by divine omnipotence forfeiting in a special way its own power over human actions.[42]

At the other extreme would be a consistent "compatibility" approach, which would hold that both good and evil, and freedom and determinism are compatible. According to this view, any act of God is justified by some good reason, even if known to and possessed by God alone; and all human acts on this view, while freely performed, are nevertheless determined by the divine will. Augustine, Aquinas, Calvin, Descartes, and Leibniz come close to this view, for they all hold that the occurrence of evil is consistent with, and therefore justifiable in terms of, the goodness of God; they attribute all good that men freely perform to God; they affirm that man's choice of evil while freely performed is not outside God's control—it is permitted by Him; though they stop short of blaming God for men's evil acts in a way that would parallel their praising him for their good ones. This approach recommends itself as the typical theist approach because it well reflects the complexity of the Biblical data on the subject, and also because it adopts a consistent semantic policy in regard to the two divine attributes of all-goodness and omnipotence.

The third possible approach would be the rare one of combining an incompatibility view on the good and evil issue with a compatibility view on the free will issue. This position characterizes Flew's discussion, except, of course, that he is not a theist. The fourth approach (and more prevalent one) combines the compatibility view of good and evil with the incompatibility view of free will and determinism. It is the view of Hick and Plantinga, each of whom holds that evil is justified and that human moral acts, in virtue of their being freely performed, lie outside the will of God to determine their moral character. I think it represents an atypical theist position going back at least as far as Pelagius.

---

[42] On an analogy with the powers of a sovereign legislature—see his discussion, *op. cit.*, pp. 210 ff.

The conflict between sceptic and theist is typically represented, then, in the contrast between the first two positions. The sceptic, according to his position, can successfully support his claim that theistic beliefs are self-contradictory if and only if he (the sceptic) furnishes those beliefs with what I have called the "incompatible" meanings for the key concepts in those beliefs. The theist, on the other hand, can successfully defend the consistency of his beliefs if and only if he provides "compatible" meanings for these concepts. When Mackie charges, then, that the theist "must be prepared to believe . . . what can be disproved from other beliefs that he also holds," he will be correct only if he is referring to theists who, like Hick and Plantinga, wish to retain the "incompatible" meanings of at least some of the concepts in these beliefs. Theists who are fully armed, however, with the "compatible" meanings of these concepts need not be prepared at all to believe what can be disproved from other beliefs that they hold. At most they must be prepared to believe what they cannot understand. For divine goodness becomes an inscrutable goodness that includes evil—indeed, given omnipotence and omniscience, it actually necessitates evil. Also divine power becomes an inscrutable power that creates, controls, and rules all things including the actions and destinies of men whose moral self-determination, in spite of this power, is altogether genuine.

But then there is some evidence that theists are prepared to believe what they do not understand. Aquinas writes, for example, concerning the names of God:

> Whatever is predicated of various things under the same name but not in the same sense is predicated equivocally. But no name belongs to God in the same sense that it belongs to creatures. . . . Therefore whatever is said of God and of creatures is predicated equivocally.[43]

Aquinas continues, of course: "Neither, on the other hand, are names applied to God and creatures in a purely equivocal sense . . . it must be said that these names are said of God and creatures in an analogous sense, that is according to proportion." The success of this attempt to moderate the equivocality of the predication lies outside the scope of the present paper; I believe it stands or falls with the comprehensibility of the divine essence, which Aquinas rejects. Then there is John Calvin, to take another example, who writes concerning divine justice:

> For the will of God is the highest rule of justice, so that what He wills must be considered just, for this very reason, because He wills it. When it is inquired, therefore, why the Lord did so, the answer must be, because He would. But if you go further, and ask why He so determined, you are in search of something greater and higher than the will of God, which can never be found. . . . The reason of the Divine justice is too high to be measured by a human standard, or comprehended by the littleness of the human mind.[44]

Now it is precisely this feature of theism—that it is a faith *lacking* understanding, or at best *seeking* understanding (but at any rate *not* possessing or demonstrating it)—that both its critics and its defenders have missed in their recent concern over the formal consistency of theism. Without an appreciation of this feature these writers are bound to obscure the real issue that divides them. Mackie distinguishes, for example, between "adequate" and "fallacious" solutions to the problem of evil. "Adequate" solutions, since they involve a forthright rejection of one of the theist beliefs, are out of the question in theistic apologetics. In what he calls the "fallacious" solutions, however, "in order to solve the problem one (or perhaps more) of its propositions is given up, but in such a way that it appears to have been retained, and can therefore be asserted without qualification in other contexts."[45]

Now if I am correct in maintaining that the formal consistency of the essential beliefs of theism can be defended simply (and only) by applying the "compatibility" logic to the concepts which these beliefs involve, then what Mackie calls (from the sceptic's viewpoint) a "fallacious" solution must be described differently. What really goes on in theistic apologetics, he should say, is that each and every proposition is *retained*, but what is *given up* is the intelligibility of its terms.

In so far as contemporary philosophical theists insist on the formal consistency of their beliefs, they stand in the tradition of rational theology.

---

[43] *Summa Theologica*, Pt. I, Q. 13, Art. 5.

[44] *Op. cit.*, Bk. III, ch. xxiii, secs. 2, 42. (See also his account of *Romans* 9 in Bk. XXX, ch. xxii, sec. 11).

[45] *Op. cit.*, p. 202.

But whether they can stand in that tradition when that tradition requires, in addition to formal consistency, the rational transparency of those beliefs—that is a deeper question which they seem to ignore. In so far as they do ignore this question they set themselves apart not only, as I hope to have indicated, from an essential ingredient in traditional theism, but also, as I suppose they well suspect, from a significant engagement with actual religious belief *and* lack of such belief in the predominantly irrational stance which both of these manifest in our time.

Grand Valley State College

Received May 7, 1966

American Philosophical Quarterly
Volume 4, Number 2, April 1967

IV. VOLUNTARY ABILITIES

TIMOTHY DUGGAN AND BERNARD GERT

WE are concerned with a particular sense of "can" that we propose to call the "can" of voluntary ability. The concept of a voluntary ability is a complex one. It includes both the concept of a physical or intellectual ability, and the concept of the ability to will to do a particular kind of action. These two concepts are themselves complex and stand in need of analysis. The terms "physical or intellectual ability," "voluntary ability," and "the ability to will" are technical terms. We believe, however, that they stand for familiar concepts which are actually employed though not fully articulated. Our discussion will be divided into three parts: the first will be devoted to an analysis of physical or intellectual abilities. The second will be devoted to an analysis of the ability to will. The third section will be devoted to an examination of certain consequences of our analysis of voluntary abilities for the traditional problems of freedom of the will and the nature of voluntary action.

I

Throughout we consistently distinguish between a *kind of action* and a *particular act*, and when we discuss physical abilities and the ability to will we are always concerned with *kinds* of action.[1] What will count as the same kind of action will depend upon which type of ability we are considering. When considering physical abilities, we say A and A' are the same kind of action if, and only if, from the fact that one has the physical ability to do A it follows that he has the physical ability to do A', and conversely.

Before we can present our analysis of the "can" of physical ability it is advisable to distinguish this sense of "can" from several other distinct, though related, senses of "can." There is, for example, that sense of "can" used in the sentence "Anyone can win the prize" in which "can" means simply that there are no restrictions put on who is allowed or permitted to win the prize; we shall not be concerned with this sense of "can" at all, nor with the sense of "can" which is used to express logical possibility. But there are several senses of "can" which are more closely related to the "can" of physical ability.

The abilities enjoyed by inanimate objects, like the ability of a machine to go a hundred miles an hour, we call *capabilities*. For these we give the following analysis: "S can do x" means: Given reasonable opportunity, S will do x a reasonable number of times. There is no mention of the order in which successes and failures must occur, provided that there is a reasonable number of successes. But there are some abilities of inanimate objects, such as the ability of titanium to withstand extremely high temperatures without melting, to which this analysis of capabilities does not apply. When we say, "This piece of titanium can withstand temperatures of over 1500 degrees," we do not mean given reasonable opportunity it will do so a reasonable number of times. For of this sort of ability, which we call a *capacity*, there is no point in talking of a *reasonable* opportunity, and even more important, when put to the test, a thing must not fail to manifest its capacities even once. To say that this piece of titanium can withstand temperatures of over 1500 degrees is to say that it will withstand such temperatures *whenever subjected to them*. If, on one occasion it does not withstand such a temperature, it can no longer be said to have that capacity. An automobile, on the other hand, may have the capability of going a hundred miles an hour, even if one day given reasonable opportunity it does not do so.

We are now in a position to distinguish between three kinds of abilities—capacities, capabilities, and physical abilities—and thus between three significantly different senses of "can" or "is able to."[2] Consider the following formulas:

---

[1] For the sake of brevity, in what follows we shall speak simply of physical abilities rather than physical or intellectual abilities although we intend that our analysis apply equally well to such abilities as the ability to recite a poem, solve crossword puzzles, and so on.

[2] We do not claim that this distinction accurately reflects ordinary usage. A glance at the dictionary is sufficient to show that "capacity," "capability," and "ability" are often used interchangeably. Our claim is simply that in philosophical discourse these distinctions will prove to be useful.

(1) "*S* can do *x*," means "*S* will never fail to do *x*." This holds of such abilities as the ability of titanium to resist high temperatures, the ability of a sheet of plate glass to flex without shattering, etc. These kinds of abilities we call *capacities*.

(2) "*S* can do *x*," means: Simply given reasonable opportunity *S* will do *x* a reasonable number of times. This formula holds for such abilities as that of the refrigerator to make ice cubes (the abilities that machines have) or the ability of a person to digest fatty foods (*some* of the abilities that persons have). These kinds of abilities we call *capabilities*.

What we call physical abilities do not fit either of the above analyses. For physical abilities, "*S* can do *x*" does not entail "*S* will never fail to do *x*." Further, if given reasonable opportunity *S* fails to do *x* a reasonable number of times, it does not follow that *S* cannot do *x*. A person may have the physical ability to run a mile in four minutes, have ample opportunity, and yet not do so a reasonable number of times simply because he does not will to do so.

"Choose," "try," "intend," "attempt," "decide," "resolve," etc., express related though different ideas. We make use of the term "will" in our analysis since it has a less circumscribed meaning than "choose," "try," "intend," etc. For example, it is a matter of debate whether or not a man who hands over his purse when threatened with physical violence can be said to choose to do so. But, as we use "will" it is correct to say that he wills to do so. In our use of "will" there is no suggestion of one's *wanting* to do the action in question, only that one did it intentionally. However, "*S* intends to do *x*" does not entail that *S* wills to do *x* because a person may intend to do this or that but when the time comes he neither does it nor does he try to do it, perhaps simply because he has changed his mind. In our sense of "will" there is no temporal gap between one's willing to do something and doing or trying to do it. If there is such a gap between "making up one's mind" to do something and doing or trying to do it, then this would be a case of intending or resolving rather than willing. The following relations hold between willing to do an action and doing it intentionally or trying to do it.

(1) If *S* tries (is trying) to do *x*, then *S* wills (is willing) to do *x*.

(2) If *S* does (is doing) *x* intentionally, then *S* wills (is willing) to do *x*.

(3) "*S* wills (is willing) to do *x*" if and only if "*S* tries (is trying) to do *x*, or *S* does (is doing) *x* intentionally."[3]

An analysis of the "can" of physical abilities must allow for the possibility that *S can* do *x* even though when given reasonable opportunity he does not do *x*. Consequently, we add to our analysis the phrase "if *S* wills to do *x*." But in order that our analysis not be too restrictive we add the phrase "But willing to do *x* is not a necessary condition for doing *x*." Thus, our analysis of physical ability does not exclude those kinds of actions which sometimes are done unintentionally. Without this addition our analysis would exclude everything that we wish to call physical abilities, e.g., the ability to walk or to talk. For one can do these things without willing to do them, e.g., in one's sleep or out of habit. However, this addition *does* rule out as physical abilities those for the exercise of which willing to do so is a logically necessary condition for doing so. Since one cannot commit murder unintentionally, the ability to commit murder is not a physical ability for willing to do so is a necessary condition for doing so. Although it sounds paradoxical to say that the ability to commit murder is not a physical ability, the paradox evaporates when it is realized that committing murder necessarily involves the exercise of physical abilities, i.e., involves doing kinds of actions which *might* be done without one's willing to do them. These same considerations apply to the ability to commit suicide and all other abilities for the exercise of which willing to do so is a necessary condition for doing so. To say of a person that he is capable of committing murder or suicide is not to say of him that he has certain physical abilities but to say something about his *character*. And in general, statements about those abilities for which willing to exercise them is a necessary condition for doing so, will be found to be statements about character.

Our proposed analysis of the "can" of physical ability is: "*S* can do *x*" means: if *S* wills to do *x* then simply given a reasonable opportunity *S* will do *x* a reasonable number of times. But willing to do *x* is not a necessary condition for doing *x*.

We now turn to objections to the analysis as a whole.

---

[3] This is not intended to be a philosophical analysis of willing, though we do provide an analysis of the *ability* to will, and consequently we are not subject to the criticism that we provide an analysis which makes essential use of the problematic notions of trying and intending. Whatever these might turn out to be, our account of willing would remain the same.

Objection 1. The analysis is too vague, i.e., the term "reasonable" which occurs several times does not admit of precise definition, whereas the statement "he can do *x*" seems to be perfectly precise.

Reply 1. This vagueness is a valuable feature of our analysis, for "he can do *x*" as expressing physical ability is only superficially precise. This is apparent from the fact that equally well-informed observers, including the person claiming (or claimed) to have the ability, will differ on what shows that he can or cannot do *x*. Analyses of "*S* can do *x*" as "*S* will do *x* if he chooses" are inadequate just because they are too precise. If I have a reasonable opportunity, it may be, and often is, that even though I fail in one or more attempts, later attempts, if successful, do indeed show that I had the ability.[4] For example, there are a number of trials or opportunities for one to demonstrate his ability to high jump, pole vault, recite a poem from memory, etc. What counts as a reasonable number of times depends on the ability claimed. If I claim to be able to swim the English Channel, one success is enough to prove me right, though I may have failed ten times previously. But if I claim to have the ability to sink short putts, one success out of ten attempts or even one out of one is not enough to show that I can. Thus from the fact that a person does something it does not follow that he has the ability to do it. But if given reasonable opportunity, I will to do *x* and consistently fail to do it, then it would be correct to say that I cannot do it. Yet, even if I fail in all my attempts, it would be a mistake to conclude that I cannot do *x* if I never, in fact, have a reasonable opportunity. And, of course, what counts as a reasonable opportunity will sometimes be an undecidable issue.

Objection 2. Granted that what will count as a reasonable number of times depends on the ability in question, even when we specify the ability, often there is no specific number of times which is *the* reasonable number of times. Yet according to the analysis any person who performs a reasonable number of times must be said to have the ability in question. Thus, there is no distinction between two people, both of whom perform a reasonable number of times, though one is more frequently successful than the other. That is, the analysis fails to account for different degrees of ability and thus for such expressions as "*S* can do *x* better than *J*."

Reply 2. On the contrary, a valuable feature of our analysis is that it does allow for degrees of ability. It is precisely because we do not equate ability with *consistent* success but only with *reasonable* success, that we can distinguish between different degrees of ability. If both Jones and Smith do *x* a reasonable number of times given the same number of attempts, and if Smith is successful a significantly greater number of times, we say that Smith can do *x* better than Jones. This, of course, is not a complete account of what is meant by "*S* can do *x* better than *J*" since other factors must be taken into consideration, e.g., the amount of time it takes for *S* to do *x* as compared with *J*, the effort involved, etc.

Objection 3. If someone never wills to do *x*, then we can never know if he would have done it a reasonable number of times. And hence, according to the analysis we can only know that a person has an ability if he wills to exercise it. But, surely we do know that, e.g., most people have the ability to touch their right knee with their left elbow though they may never have willed to do so.

Reply 3. Our analysis includes the conditional form —if *A* and *B*, then *C*, thus when one or both of the antecedents are false, the result is a counterfactual conditional. The objection presupposes that whereas we might know that someone has an ability which he has never exercised, the corresponding counterfactual yielded by our analysis could not be known to be true. But it is simply false that we can never know the truth of counterfactuals. For example, there is no doubt that statements such as "if the vase had been dropped it would have shattered" are sometimes true, and known to be true. For every case in which a statement attributing an unexercised ability to a person is true and known to be true, the corresponding counterfactual statement yielded by our analysis will be true and known to be true. In trivial cases where there is no doubt about the ability, there will be no doubt about the corresponding counterfactual. In those cases in which the counterfactual is doubtful there will be a similar uncertainty about the ability.[5]

---

[4] But not always. In the meantime I may have acquired the ability. In our analysis we say "*simply* given reasonable opportunity" in order to exclude this kind of case. Abilities, like capabilities, can be developed, lost, and regained. Our analysis concerns a person's abilities given the condition he is in at the time, not the condition he was or will be in. Thus we distinguish physical abilities from what we call "natural abilities." "*S* has the natural ability to do *x*" means: With natural development and/or a reasonable amount of training, *S* will acquire the physical ability to do *x*. The less training required the greater the natural ability.

Objection 4. The proposed analysis is intended to account for the "can" of physical ability. But if we apply it to the case of a physically fit person suffering from extreme claustrophobia, we might be forced to say that he can walk into a closet and shut the door even though it is obvious that he cannot do so.

Reply 4. This objection is the result of confusing the "can" of physical ability and what we call the "can" of voluntary ability. Simply because a person has the physical ability to do a kind of action, it doesn't follow that he has the voluntary ability to do that kind of action. For in addition to the physical ability to do a kind of action, one must also have the ability to will to do it in order to have the voluntary ability to do it.

II

A voluntary ability consists of both a physical ability and the ability to will. Our concern up to now has been with the "can" of physical ability. But, to provide an analysis of "*S* can do *x*" where this expression refers to a voluntary ability, we must now provide an analysis of the ability to will. To say that *S* can will to do *x* (a kind of action) is to say that it is possible to describe some state of affairs such that *S* would believe there were reasonable incentives for doing *x* and believing this, would will to do *x*; and it is possible to describe some state of affairs such that *S* would believe there were reasonable incentives for not doing *x* and believing this, would will not to do *x*.[6]

One might be tempted to substitute for our analysis of the ability to will the following less complex formula: It is possible to describe some reasonable incentives for doing *x* such that *if S* believed they were present he would will to do *x*, etc. However, this would be an inadequate substitute because it would allow that one might have the ability to will even though he lacks an essential element, viz., it might be that *S* cannot believe that there are any reasonable incentives for doing *x*. Thus one would have to add to the above, "*S can* believe that there are reasonable incentives for doing *x*." It would then be necessary to analyze this in the following way: It is possible to describe some state of affairs such that *S* would believe there were reasonable incentives for doing *x*, etc. We believe that this temptation should be resisted because in addition to our view that one should not multiply analyses beyond necessity, separating the analysis of "can believe . . ." from the analysis as a whole might suggest that a person could have the ability to will even though he could not have the necessary beliefs about reasonable incentives.

Incentives for willing to do *x* are a subclass of the envisaged consequences of doing *x*, where (1) *x* is a kind of action that one might have the physical ability to do and (2) the description of *x* contains no incentive either for doing or refraining from doing the action. They are that subclass which a reasonable man would regard as reasons for willing to do *x*. An unreasonable incentive for willing to do *x* is one which it would be unreasonable to expect any reasonable man to resist. A reasonable incentive for willing to do *x* is an incentive for willing to do *x* which is not unreasonable. Of course, what counts as an unreasonable incentive will depend on the kind of action in question and the situation of the agent.

From the analysis it is clear that it will be much easier to show that one *has* the ability to will than to show that one *lacks* it. Since to show that one has the ability, one need only will to do *x* when one believes that there are reasonable incentives for doing it, and will not to do *x* when one believes that there are reasonable incentives for not doing it. However, to show that one lacks the ability to will to do *x* we would have to establish either that it is impossible to describe any state of affairs such that *S* would believe there were reasonable incentives for doing *x* and believing this would will to do *x*, or that it is impossible to describe any state of affairs such that *S* would believe there were reasonable incentives for not doing *x* and believing this would will not to do *x*.

Obviously, it is extremely difficult to establish either of these alternatives. Nonetheless, we can have evidence, sometimes overwhelming evidence, that one lacks the ability to will to do *x*. We can have this evidence because we can present *S* with numerous and varied reasonable incentives for willing to do *x*. When trying to determine whether or not S has the ability to will, it is not necessary that the incentives actually be present, but only that the person believe that they are. Thus, though there are often a limited number and variety of reasonable incentives that can actually be presented, there is virtually no limit on the number and variety of reasonable incentives that S can be induced to believe are present, by hypnosis or deceit.

---

[5] A parallel objection and reply could be made in the case where there is never a reasonable opportunity.

[6] Note that we say "*S* would will not to do *x*" rather than "*S* would not will to do *x*" (which does not seem to be strong enough) or "*S* would will to do not—*x*" (which seems unintelligible unless it simply means "*S* would will to refrain from doing *x*" which is equivalent to what we say).

In determining whether one has or lacks the ability to will to do x, it is important to present a number of different *kinds* of incentive. That no amount of money would induce S to will to do x does not show that he lacks the ability to will to do it. For he may will to do it quite readily when presented with a different kind of reasonable incentive, e.g., the avoidance of pain or shame for himself or others. We do not hold that a man who has the ability to will to do x will will to do x whenever he believes that a reasonable incentive for willing to do x is present. Thus, a man may have the ability to will to do x, believe that reasonable incentives for willing to do x are present, and yet not will to do x. But, if he fails to will to do x, no matter what reasonable incentives he believes to be present, then he would conclude that he lacked the ability to will to do x.

This can be seen from the following case. Imagine that S has been hypnotized so that he believes everything that he is told by the hypnotist. Thus, he believes that such and such will be a consequence of his willing to do x, when told this by the hypnotist. Let us suppose that he is told that he will receive a certain sum of money for willing to do x, and yet fails to will to do so. Then he is told that he will avoid pain if he wills to do x, and still fails to do so. Next he is told that others will avoid suffering if he wills to do x, and still fails to do so. If after having been presented with numerous and varied reasonable incentives S still does not will to do x, then we have conclusive evidence that he does not have the ability to will to do x. We do not suggest the use of hypnosis to establish the *presence* of an ability, but only the *absence* of an ability. We use this example simply because hypnosis is such a simple and effective way to deceive someone.

Further, even if a man insists that he would will to do x if there were reasonable incentives for doing it and yet it is impossible to describe *any* state of affairs which he would regard as providing reasonable incentives for doing x, then he lacks the ability to will to do x. Imagine the following case: A lukewarm pacifist claims that *if* there were reasonable incentives for taking a human life then he would be prepared to do so, but that it is impossible to describe states of affairs which he would regard as providing adequate grounds for killing a human being. That is, he does not regard the preservation of society, the preservation of numerous innocent lives, much less the acquisition of wealth and power for himself, etc., as adequate grounds for killing. Such a person, we conclude, lacks the ability to will to take a human life.

When discussing the ability to will our basic analysis of *same kind of action* is in terms of the *lack* of the ability to will. For, normally when people are concerned with the ability to will, they are concerned with the kinds of action that one lacks the ability to will. That is, they group together the abnormal cases not the normal ones. People who suffer from compulsions as well as those who suffer from phobias lack the ability to will to do certain kinds of actions. The compulsive hand washer *lacks* the ability to will to wash his hands precisely because there are no reasonable incentives for which he would refrain from doing so. Here our analysis of same kind of action is: A and A' are the same kind of action for S if they share a feature such that S *lacks* the ability to will to do any action having that feature. What will be the same kind of action for one person may not be the same kind of action for another. However, frequently the same kind of action for one person will be the same kind of action for many other persons, e.g., for all extreme claustrophobiacs, A and A' will be the same kind of action provided that they share the feature of entering a small confined place.

Even though we naturally talk about kinds of action that one *lacks* the ability to will, it will prove useful to offer an analysis of what counts as the same kind of action that one *has* the ability to will. Thus, we say: A and A' are the same kind of action that S has the ability to will, if and only if A and A' are the same kind of action with regard to physical abilities (see above) *and* neither A nor A' are a kind of action that S lacks the ability to will. Admittedly a consequence of this analysis is grouping together as the same kind of action, actions which are radically different; e.g., for someone with the ability to will to strangle his wife, strangling his wife will be the same kind of action (with respect to his having the ability to will) as tightening his fingers with a certain amount of force around an object having a certain diameter, resiliency, etc. However, we think that the usefulness of being able to talk about kinds of actions one *has* the ability to will should be apparent from the following: every kind of action that one has the ability to will to do involves a kind of action completely describable as a kind of action that one has the physical ability to do. That is, the ability to will to strangle one's wife involves the ability to will to tighten one's fingers with a certain amount of force around a physical object of a certain diameter and resiliency. Some of the kinds of actions that one has (or lacks) the ability to will to do are completely describable as kinds of actions one has the physical ability to do, e.g., lift a 10-pound bag of sugar. Of course, having the ability to will to do a kind of action does not entail having the physical ability to do it.

As we use the term "physical ability," the kinds of actions that one has or lacks the physical ability to do are describable in such a way that the description, taken by itself, provides no incentives for or against doing the action. If the description of those kinds of actions which one has the ability to will is the same as a description of kinds of actions one has the physical ability to do, then taken by themselves, these descriptions also will provide no incentives for or against willing to do that kind of action. On the other hand, descriptions of those kind of actions that one may have the ability to will which are *not* identical with some description of a kind of action that one may have the physical ability to do, taken by themselves, may provide incentives either for or against willing to do that action. For example, the description "strangling one's wife" does provide a *prima facie* incentive for willing not to do that action. Descriptions of these kinds of actions can be replaced by descriptions of kinds of actions which one might have the physical ability to perform plus a description of the envisaged consequences of this exercise of one's physical ability. Therefore, by "subtracting" from the description of a kind of action that one might have the ability to will that part which describes a kind of action that one might have the physical ability to do we may be left with a description of an incentive for willing to do or willing to refrain from doing that action.

What we have in mind is illustrated by the following: "strangling one's wife" may be regarded as a description of a kind of action that one may have the ability to will to do. Let us replace this description with the following: "tightening one's fingers with a certain amount of force around an object of a certain diameter and resiliency, etc., consequently doing serious physical harm to one's wife." Now, if we subtract from the above that part dealing with physical abilities only, we have as a remainder a *prima facie* incentive for refraining from the action. We do not say that the latter description is equivalent to the former, but only that it may be a description of a kind of action which is described by the former.

For similar persons in similar situations, incentives which would be unreasonable for willing to do an action completely describable as a kind of action one may have the physical ability to do may be reasonable incentives for willing to do those actions the description of which includes a *prima facie* incentive. That is, if the description of an action includes a reasonable incentive for willing to do the action, it will be reasonable to offer much greater incentives for willing to refrain from doing the action, and conversely. It might be said that the less incentive required for differing choices, the greater is the ability to will to do that action. This sounds very plausible, in that it makes trivial kinds of actions those most subject to the will and momentous ones less so.

Just as from the fact that a person does x it does not follow that he has the physical ability to do x, so too from the fact that a person *wills* to do x it does not follow that he has the ability to will to do x. The example of the compulsive hand washer mentioned earlier is a clear case of a person who although he frequently wills to wash his hands lacks the ability to will to do so. The person suffering from extreme claustrophobia may provide another case. For such a person there are no *reasonable* incentives for which he would will to go into a small closet and shut the door. That is, he would not will to do so for a reasonable sum of money, or to develop some film, etc. We want to say that our claustrophobiac lacks the ability to will to go into the closet. And yet there might be unreasonable incentives given which he would will to do so, e.g., at the point of a sword, or if this was the only way to escape from a horde of poisonous spiders.[7] Our claustrophobiac does not have the voluntary ability to walk into a closet and shut the door even though he has the physical ability to do so, i.e., *if* he *wills* to do so then, given a reasonable opportunity he will do so a reasonable number of times. It's just that he lacks the ability to will to do so.

---

[7] Just as with capacities, capabilities, and physical abilities, the ability to will to do a kind of action may be gained or lost, e.g., when a person is changed by undergoing psychoanalytic treatment, or conversely, when he develops a phobia.

III

As we have described them, voluntary abilities are a combination of physical or intellectual abilities and the ability to will. One must have both of these in order to have a voluntary ability. Voluntary abilities just like physical abilities and the ability to will are concerned with kinds of actions rather than particular acts. Since we have, so far, been concerned only with the ability to do kinds of actions rather than particular acts, we have said nothing which *prima facie* is relevant to the problems of free will and of the nature of voluntary action. However, we do think that our analysis of voluntary ability can contribute something to the resolution of these issues, and in this section, we hope to make this evident.

We are now in a position to answer four questions, viz., what does it mean to say, (1) S can do a particular act? (2) S did a particular act voluntarily? (3) S is free to do a particular act? (4) S did a particular act freely? In answering some of these questions it is necessary to use the notion of voluntary ability so that it would be useful to present the complete analysis of voluntary ability here. "S can do x," i.e., "S has the voluntary ability to do x" means: If S wills to do x, then simply given reasonable opportunity, S will do x a reasonable number of times. But willing to do x is not a necessary condition for doing x. And S can will to do x, i.e., it is possible to describe some state of affairs such that S would believe there were reasonable incentives for doing x and believing this, would will to do x; and it is possible to describe some state of affairs such that S would believe there were reasonable incentives for not doing x and believing this, would will not to do x.

"S can do a particular act" means S has the voluntary ability to do that *kind* of action, and a reasonable opportunity to do it. "S did a particular act voluntarily" means S had the ability to will to do that kind of action, willed to do it, and did it. Thus, the answers to the first two questions can be given very simply once the concept of a voluntary ability and related concepts have been worked out. But, while these answers are easily arrived at, they are not so easily understood.

Suppose we are wondering, if here and now, S can do x. According to the answer that we have given, only two factors are relevant: (a) Does he have the voluntary ability to do that kind of action, and (b) Does he have a reasonable opportunity to exercise that ability here and now? If we answer both of those questions in the affirmative, then, according to our analysis, he can do the particular act under consideration.

This, so far, is not surprising, or troublesome; it seems consistent with what we ordinarily think about the matter. But it does become somewhat puzzling in those cases in which the person who has the voluntary ability and a reasonable opportunity does in fact will to do the action in question and yet fails to do it. It may be objected that according to our analysis it is possible that all the conditions for a particular act being done are satisfied and yet the agent fails to perform as expected? This objection, however, assumes that failure to do an act under the conditions envisaged entails that one could not have done that particular act. It seems to us that this is false. We do allow that it is possible that a person can do a particular act, and yet not do it even though he has a reasonable opportunity and he wills to do it. And our analysis allows for this; for in establishing a voluntary ability we do not demand consistent success, only reasonable success. Thus, it is perfectly consistent, according to our analysis to conclude that a man can do a particular act, and to stick to this even when he has a reasonable opportunity, wills to do it and fails.

Though there may be some sense of "He can do this," when "this" refers to a particular act, in which all that is involved is his doing it here and now, this seems to us at best a colloquialism with little or no philosophical interest. Perhaps this may be seen from the following: Suppose someone says he can sink a very tricky putt, does so, and says with glee "You see I told you so," and his partner replies "Just a matter of luck." There is a genuine dispute here, one that could be settled by further facts, not merely by one person repeating that he had done it. But if one takes "He can do this" merely to be a prediction that he will do this, then there could be no dispute, and one is not talking about abilities at all.

The answer to the second question—"What does it mean to say that S did a particular act voluntarily?"—also raises some interesting issues. There is certainly a difference between a man doing something intentionally and his doing it voluntarily, and yet it is not that doing something intentionally and doing it voluntarily are *doing* different kinds of things. To do something voluntarily it is necessary that one does it intentionally, but this is not sufficient. A person may intentionally do a number of things while under the influence of drugs or suffering from some neurosis or psychosis. Yet we do not want to say that these things are done voluntarily. According to our analysis, the difference between an intentional act and a voluntary one lies not in what is done on that particular occasion, but on whether the person in question had the ability to will to do that kind of action. Thus, an act is done voluntarily if the agent has the ability to will to do that kind of action and does in fact intentionally do it on a particular occasion.

Since having the ability to will to do a kind of action includes the ability to will *not* to do that kind of action, this analysis distinguishes clearly between intentional and voluntary acts. For an intentional act done by one who lacks the ability to will to do that kind of action (a paradoxical sounding, but perfectly intelligible case) is not a voluntary act, but a compulsive one or something like that. Our distinction between a voluntary act and an intentional one does not depend upon a fictitious mental occurrence, but on the state of the agent. We may know that a person did something intentionally and yet not know whether it was done voluntarily simply because we do not know enough about the mental state of the person in question, i.e., we do not know if he had the ability to will to do the kind of action involved (in spite of the fact that he did indeed will to do it).

From what we have said so far, it is possible that one does not have the voluntary ability to do a particular act and yet do it voluntarily. This seeming paradox arises from the fact that for *S* to have the voluntary ability to do a particular act he must have both a physical or intellectual ability and the ability to will, whereas for *S* to have done the act voluntarily all that is required is that *S* had the ability to will to do that kind of action and intentionally did it. Consequently, one may lack the physical ability to do a kind of action, but have the ability to will to do it, and in fact, will to do it and do it on a particular occasion. The agent would have done the act voluntarily even though he did not have the voluntary ability to do it. What we have in mind is that people exceed their abilities on occasion. In a moment of great stress it is perfectly possible for a person voluntarily to do something while it would also be correct to say he lacks the voluntary ability to do that kind of thing, e.g., a woman lifts an automobile thus freeing her child trapped beneath.

Our answer to the question, "What does it mean to say that *S* is free to do a particular act?" makes use of the account given of "*S* can do a particular act." A person may have the voluntary ability to do a kind of action and a reasonable opportunity but at the same time be under duress such that his will is unduly influenced. We say, therefore, that a person is free to do a particular act if, and only if, he has the voluntary ability to do that kind of action, a reasonable opportunity, and there are no unreasonable incentives influencing his will either to do or to refrain from doing that act.

A question which we have not yet answered is: What does it mean to say that *S* did a particular act freely? Here we make use of our answer to the second question: "What does it mean to say that *S* did a particular act voluntarily?" For an act to be done freely it must be done voluntarily, but this is not enough. According to our analysis an act done voluntarily must be one of a kind that a person has the ability to will to do and that he does will to do. But, although he has the ability to will to do that kind of action, and does in fact will to do it on this particular occasion, thus doing it voluntarily, it may be that on this particular occasion he willed to do it because he was subjected to unreasonable incentives. Here we should not want to say that his act was done freely. Thus our analysis of a free act, or an act freely done, is one which a person does voluntarily, i.e., he does it intentionally and has the ability to will to do it, and there were no unreasonable incentives successful in leading him to do it.

Consequently, we make a sharp distinction between those cases in which one is subjected to unreasonable pressures or incentives which force him to will to do an act which he has the ability to will to do, and those cases in which he does not have the ability to will at all. The latter we do not even call voluntary acts, the former we call voluntary but not free. Thus, we make a distinction between normal people caught in extraordinary circumstances who will to do an act because they were subjected to unreasonable incentives and the mentally ill who intentionally do an act because they do not have the ability to will otherwise. Of course, this distinction is not absolutely sharp. There are cases in which we may not be sure whether one has been subjected to unreasonable incentives or whether one's ability to will has been taken away, e.g., brainwashing.

The importance of distinguishing between voluntary acts and free ones lies primarily in questions of assigning responsibility. An act which is not a voluntary one is not one for which the agent is responsible. But an act which is not free because of the presence of successful unreasonable incentives may still be one for which we might assign responsibility. As long as we think someone's ability to will is intact, which it is in a voluntary but unfree act, we think it may be appropriate to assign responsibility. We may hold that no matter how unreasonable the incentives, some acts should not be done. We do in fact distinguish between those who were unable to do other than they did, and those who though they had the ability to do otherwise were subjected to unreasonable incentives. The latter is a matter of degree, the former an absolute. In the latter case when the person under consideration is at the time of his action thought to have the ability to will, then no matter what the circumstances were, he *may* be held responsible to some degree for his voluntarily doing the act. Voluntary acts are always susceptible to praise and blame, though to say that they were *not* freely done is to admit that the praise or the blame is to be softened, perhaps even completely withheld.

There are, of course, problems in deciding what are reasonable and what unreasonable incentives, but these problems cannot be solved in the abstract, rather they must be related to individual cases. This notion of reasonable incentives is important, for not only does the presence of successful unreasonable incentives make an act not free, or perhaps better, less free, but also the notion of reasonable incentives is important in establishing whether someone has the ability to will to do a kind of action or not.

In ordinary discourse the term "voluntary" as applied to actions has a number of senses which though related are nonetheless, distinct. Of these three are central (1) a voluntary act is simply an intentional (as opposed to an accidental) act; (2) a voluntary act is a free (as opposed to a constrained) act; (3) a voluntary act is one which proceeds from the will of an agent who has what we have described as the ability to will to do that kind of action.

The terms "intentional act" and "free act" seem to be appropriate for the first and second categories. We reserve the term "voluntary act" for the third category. We do not claim that these distinctions accurately reflect any consistent pattern of ordinary usage. We simply claim that the conceptual distinctions developed in this paper will prove useful in philosophical discussion of the problem of free will and the nature of voluntary action.

Dartmouth College

Received April 19, 1966

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 2, April 1967

V. DEONTIC LOGICS

G. H. VON WRIGHT

I

THE existing systems of (propositional) deontic logic can conveniently be divided into two groups. I shall call them the *monadic* systems and the *dyadic* systems.

The formal language of a (propositional) deontic logic has the following *vocabulary*:

a. Variables p, q, r, . . . an unlimited supply, representing some "proposition-like" entities.
b. Truth-connectives $\sim$, &, v, $\rightarrow$, and $\leftrightarrow$ for negation, conjunction, disjunction, material implication, and material equivalence respectively.
c. Two deontic operators. In the monadic logics these are P and O. In the dyadic logics they are $P(-/-)$ and $O(-/-)$.
d. Brackets.
e. An auxiliary constant t which stands short for an arbitrary tautology of propositional logic (PL), e.g., for $p\text{v}{\sim}p$.

To save brackets we adopt the following convention concerning the "binding force" of the connectives: $\sim$ is the weakest and then follow in order of increasing strength &, v, $\rightarrow$, and $\leftrightarrow$. By virtue of this convention we can, e.g., for $(((\sim(p)\&q)\text{v}r)\rightarrow s)\leftrightarrow u$ write simply ${\sim}p\&q\text{v}r\rightarrow s\leftrightarrow u$. We also have the convention that when O and P are followed by a variable, the variable need not be in brackets. Thus, e.g., for $P(p)$ we write simply Pp.

The *expressions* (well-formed formulae) of a deontic logic we (here) define as follows:

a. An expression of the form "$P(-)$" or "$O(-)$" or "$P(-/-)$" or "$O(-/-)$" is well-formed when the place of "–" is taken by an expression of PL.
b. Truth-functional compounds of well-formed expressions are well-formed.

When the expressions of a deontic calculus are defined in this way, "mixed" expressions, i.e., truth-functional compounds of expressions of PL and of expressions of the deontic calculus, do not count as well-formed. Nor do expressions count as well-formed which *iterate* a deontic operator.

These restrictions on the well-formed formulae can be removed. We shall not, however, in this paper consider the more general systems of deontic logic which admit "mixed" expressions and iteration of deontic operators.

'P' may be read as "it is permitted that" and 'O' may be read as "it is obligatory that" or, alternatively, "it is forbidden that not." Another way of reading the operators would be "it is permitted to see to it that" and "it is obligatory to see to it that" or, alternatively, "it is forbidden to see to it that not." Under the second reading, the permitted and obligatory things will be those *actions* which result in the *states of affairs* described by the expressions of PL within the scope of the deontic operator.

There is a monadic deontic logic with the following axioms:[1]

A0M. A set of axioms of PL (with well-formed expressions substituted for the variables).
A1M. $Pp \leftrightarrow {\sim}O{\sim}p$.
A2M. $P(p\text{v}q) \leftrightarrow Pp \text{ v } Pq$.
A3M. $Pp \text{ v } P{\sim}p$.

The rules of inference of the calculus are:

R1. Substitution (of formulae of PL for the variables).
R2. Detachment (*modus ponens*).
R3. A Rule of Extensionality to the effect that provably equivalent formulae of PL are intersubstitutable in the wffs of the deontic calculus.

A dyadic deontic logic which, for reasons to be given later, may be called the "counterpart" of the above calculus has the same rules of inference and the following axioms:[2]

---

[1] This is, from the formal point of view, substantially the same as the system of my paper "Deontic Logic," *Mind*, vol. 60 (1951), pp. 1–15. What is here A1M was there regarded as a *definition* of the one deontic operator in the terms of the other (and negation).

A0D. A set of axioms of PL (with well-formed expressions substituted for the variables).

A1D. $P(p/q) \leftrightarrow \sim O(\sim p/q)$.

A2D. $P(p \vee q/r \vee s) \leftrightarrow P(p/r) \vee P(p/s) \vee P(q/r) \vee P(q/s)$.

A3D. $P(p/q) \vee P(p/\sim q) \vee P(\sim p/q) \vee P(\sim p/\sim q)$.

By virtue of the distribution-axioms A2, we could also write A3M in the form Pt and A3D in the form $P(t/t)$.

II

It seems that in all known systems of deontic logic one can prove certain formulae which will strike one as counterintuitive or somehow "paradoxical." I shall here mention a few such paradoxes. They all arise in the monadic system which was outlined in the preceding section—and reappear, *mutatis mutandis*, in the dyadic system.

We easily prove that $Op \rightarrow O(p \vee q)$. This says that an obligation to see to it that (it is the case that) p entails an obligation to see to it that (it is the case that) p *or* q. So, for example, if I ought to mail a letter, I also ought either to mail *or* to burn it. This seems odd. The oddity was first pointed out by the Danish jurist-philosopher Alf Ross in a paper which may be called a trenchant criticism of the very idea of a "deontic logic."[3] I shall refer to this oddity under the name of Ross's Paradox.

This paradox has often been met by the following argument (or by some argument similar to it): If I see to it that p, and thereby fulfill the first of the two obligations above, then I also, by the laws of ordinary logic, see to it that it is the case that p *or* q. But from this it cannot, by the laws of any logic ("ordinary" or "deontic") be concluded that it were obligatory or even permitted for me to see to it that q. To say that Op entails $O(p \vee q)$ is really no more paradoxical than to say that p entails $p \vee q$. Beginners in logic may find the last a little bit hard to swallow too, but they will soon get over the difficulty.

This argument has seemed satisfactory to many people. I have tried myself to be pleased with it, but never quite successfully. I always felt that there was more to Ross's Paradox than can be met by the above argument. I hope to be able to show that this feeling is justified. (But I shall not follow Ross in concluding that deontic logic is therefore impossible.)

We also prove that $\sim Pp \rightarrow \sim P(p \& q)$. This says that if it is forbidden (not permitted) that it be the case that p, then it is also forbidden (not permitted) that it be the case that p *and* something else. This seems odd, particularly if q is something which we ought to see to that is the case, when p is the case. This last oddity is sometimes called the Paradox of the Good Samaritan.

A further thing which we can prove is that $\sim Pp \rightarrow O(p \rightarrow q)$. This says that, if it is forbidden that it be the case that p, then it is our duty to see to it that it is the case that q, *if* it is the case that p—and this no matter what state of affairs q is. This oddity is known as a Paradox of Derived Obligation or of Commitment.[4]

It is of some interest to note that, if we accept the "interdefinability" of the deontic operators as laid down in A1M, then the three formulae "$Op \rightarrow O(p \vee q)$" and "$\sim Pp \rightarrow \sim P(p \& q)$" and "$\sim Pp \rightarrow O(p \rightarrow q)$" are but variations of one and the same logical structure.

Finally, I mention the following perplexity. On the ordinary understanding of "it is permitted that," $P(p \vee q)$ seems to entail $Pp \& Pq$. If I say to somebody "you may work or relax" I normally mean that the person addressed has my permission to work and also my permission to relax. But, in what may be called the "standard" systems of deontic logic, $P(p \vee q)$ does *not* entail $Pp \& Pq$. If it did, we would moreover get strange consequences. We could then, in the system presented in Section I, easily derive as a theorem that $Op \rightarrow Pq$, which says that if we ought to see to it that p, then we are permitted to see to it that q, no matter what this second state of affairs is. And from this result we could derive a contradiction in the system.

---

[2] The variety of dyadic systems seems to be greater than that of monadic systems. The, as far as I know, historically first dyadic system was proposed in my "A Note on Deontic Logic and Derived Obligation," *Mind*, vol. 65 (1956), pp. 507–509. A different system is due to N. Rescher, "An Axiom System for Deontic Logic," *Philosophical Studies*, vol. 9 (1958), pp. 24–30. The system A0D–A3D which is given here was first presented by me in "A New System of Deontic Logic," *The Danish Yearbook of Philosophy*, vol. 1 (1964) and "A Correction to a New System of Deontic Logic," *ibid.* vol. 2 (1965). In these papers, the axioms were stated in terms of the obligation operator "O(–/–)" and A2D was split up in two principles, *viz.*, $O(p \& q/r) \longleftrightarrow O(p/r) \& O(q/r)$ and $O(p/q \vee r) \longleftrightarrow O(p/q) \& O(p/r)$. The truth of A1D was regarded as definitional.

[3] Alf Ross, "Imperatives and Logic," *Theoria*, vol. 7 (1941), pp. 53–71.

[4] See A. N. Prior, "The Paradoxes of Derived Obligation," *Mind*, vol. 63 (1954), pp. 64–65, and my reply "A Note on Deontic Logic and Derived Obligation," *Mind*, vol. 65 (1956), pp. 507–509. A challenging expression of dissatisfaction with the situation in deontic logic with regard to this paradox is R. M. Chisholm, "Contrary-to-Duty Imperatives and Deontic Logic," *Analysis*, vol. 24 (1963), pp. 33–36.

One could meet this paradox by saying that a disjunctive permission is *compatible* with the permissibility of both the disjuncts, but that it does not *entail* their permissibility. One could then admit that *normally* $P(p \vee q)$ goes together with $Pp \& Pq$ but that this is no (logical) *necessity*. I have tried to meet the paradox in this way in past writings of mine.[5] But I no longer think the defense entirely successful.

III

The occurrence of "paradoxes" in intuitively seemingly sound systems of monadic deontic logic, and the vagueness and conflicting nature of the intuitions upon which the dyadic systems build[6] make urgent the need of giving to all those systems a sound semantic foundation. In a recent paper,[7] Nicholas Rescher has made an important contribution toward satisfying this need.

Rescher's contribution immediately concerns the dyadic systems. He proposes four different interpretations of the notion of a conditional permission, $P(p/q)$, in the terms of relations between sets of *state-descriptions* or "possible worlds." I shall here exploit Rescher's idea for showing how a variety of (monadic and dyadic) systems of deontic logic may become generated from a common pattern. The systems, thus generated, are decidable and semantically complete with regard to a specified criterion of logical truth.

I introduce the following piece of new terminology: An expression of the form "$P(s/s')$," where 's' and 's'' are state-descriptions, may be read "in the possible world s' the possible world s is permitted (as an alternative world to s')." Similarly, an expression of the form "$O(s/s')$" may be read "in the possible world s' the possible world s is obligatory (as an alternative world to s')" or also as "in the possible world s' any possible world other than s is forbidden (as an alternative world to s')." The expression "$P(s/s')$" thus says that a world of the description 's'' *may* become changed to a world of the description 's'. The expression "$O(s/s')$" says that a world of the description 's'' *ought to* be thus changed, i.e., *must not* be changed to anything but this (nor left unchanged).

IV

I define four concepts of conditional permission as follows:[8]

"$P_1(p/q)$" will mean that in *some* possible world in which it is true that q *some* possible world is permitted in which it is true that p.

"$P_2(p/q)$" will mean that in *all* possible worlds in which it is true that q *some* possible world is permitted in which it is true that p.

"$P_3(p/q)$" will mean that in *some* possible world in which it is true that q *all* possible worlds are permitted in which it is true that p.

"$P_4(p/q)$" will mean that in *all* possible worlds in which it is true that q *all* possible worlds are permitted in which it is true that p.

To the four concepts of conditional permission there answer four concepts of conditional obligation, "$O_1(-/-)$", "$O_2(-/-)$", "$O_3(-/-)$", and "$O_4(-/-)$":

"$O_1(p/q)$" which will mean that in *all* possible worlds in which it is true that q *all* possible worlds are forbidden in which it is *not* true that p;

"$O_2(p/q)$" which will mean that in *some* possible world in which it is true that q *all* possible worlds are forbidden in which it is *not* true that p;

"$O_3(p/q)$" which will mean that in *all* possible worlds in which it is true that q *some* possible world is forbidden in which it is *not* true that p; and

"$O_4(p/q)$" which will mean that in *some* possible world in which it is true that q *some* possible world is forbidden in which it is *not* true that p.

We set up four dyadic calculi which we shall refer to as the *weak* P_1-, P_2-, P_3-, and P_4-calculi.

---

[5] "Deontic Logic," *Mind*, vol. 60 (1951), p. 7 n.

[6] See the exchange of thoughts between Rescher and Anderson in the papers: N. Rescher, "An Axiom System for Deontic Logic," *Philosophical Studies*, vol. 9 (1958), pp. 21–30; A. R. Anderson, "On the Logic of 'Commitment'," *Philosophical Studies*, vol. 10 (1959), pp. 23–29; N. Rescher, "Conditional Permission in Deontic Logic," *Philosophical Studies*, vol. 13 (1962), pp. 1–6; and A. R. Anderson, "Reply to Mr. Rescher," *Philosophical Studies*, vol. 14 (1963), pp. 6–8. See also J. Robison, "Further Difficulties for Conditional Permission in Deontic Logic," *Philosophical Studies*, vol. 18 (1967).

[7] N. Rescher, "Semantic Foundations for Conditional Permission," *Philosophical Studies*, vol. 18 (1967). For a semantic approach to monadic deontic logic see W. H. Hanson, "Semantics for Deontic Logic," *Logique et Analyse*, vol. 8 (1965), pp. 177–190.

[8] These four concepts of conditional permission are not the same as those of Rescher.

The vocabulary and expressions are defined as in Section I.

All four calculi have this axiom in common:

A0 DP$_i$. A set of axioms of PL (with well-formed expressions substituted for the variables).

All four calculi also have an axiom of the form:

A1 DP$_i$. $P_i(p/q) \leftrightarrow \sim O_i(\sim p/q)$, where for '$i$' is substituted '1' in the P_1-calculus, and . . . and '4' in the P_4-calculus.

Each calculus has its own characteristic *distribution axiom*:

A2 DP$_1$. $P_1(p \vee q/r \vee s) \leftrightarrow P_1(p/r) \vee P_1(p/s) \vee P_1(q/r) \vee P_1(q/s)$.

A2 DP$_2$. $P_2(p \vee q/r \vee s) \leftrightarrow (P_2(p/r) \vee P_2(q/r))\ \&\ (P_2(p/s) \vee P_2(q/s)\ )$.

A2 DP$_3$. $P_3(p \vee q/r \vee s) \leftrightarrow (P_3(p/r)\ \&\ P_3(q/r)) \vee (P_3(p/s)\ \&\ P_3(q/s)\ )$.

A2 DP$_4$. $P_4(p \vee q/r \vee s) \leftrightarrow P_4(p/r)\ \&\ P_4(p/s)\ \&\ P_4(q/r)\ \&\ P_4(q/s)$.

The rules of inference of all four calculi are the rules R1–R3 (Section I).

Comment on the truth of the axioms: A1 may be regarded as following from a proposed definition of the one deontic operator in the terms of the other (and negation). The distribution axioms A2 again are the syntactical equivalents of truths which in the semantics of the calculi follow directly from the definitions of the various permission-concepts.

The expressions of each calculus have what I propose to call a P_i-normal form. This we obtain as follows: If the expression is a truth-functional compound which contains constituents of the form "$O_i(-/-)$" we replace these constituents by constituents of the form "$P_i(-/-)$" by virtue of A1. Thereupon we replace, by virtue of R3, the expressions to the left and to the right of '/' in the constituents by their disjunctive normal forms in terms of *all* the variables which occur in the whole expression.[9] Finally, we distribute the constituents by virtue of the appropriate A2 as far as possible. The result of the distribution is the P_i-normal form of the expression.

We adopt the following *criterion of logical truth* in a weak dyadic deontic logic: *An expression of a P_i-calculus is logically true, if and only if, its P_i-normal form is a tautology of propositional logic (PL).*

By transforming a given expression into its normal form we can immediately see whether it is a logical truth or not. It is obvious that all formulae which are, on the set criterion, logically true in the P_i-calculus are also provable in the P_i-calculus (by virtue of A0) and that all formulae which are provable in the P_i-calculus are, on the set criterion, logically true. Thus, relative to the truth-criterion, the calculi are semantically complete.

V

To the axioms of each one of the weak calculi which we described in the preceding section we add a fourth axiom:

A3 DP$_i$. $P_i(t/t)$. ($1 < i < 4$.)

The four calculi which we then get I shall call the *strong* (dyadic) P_1-, P_2-, P_3-, and P_4-calculus. A3 I shall call the Principle of Permission.[10]

"$P_1(t/t)$" says that in *some* possible world *some* possible world is permitted.

"$P_2(t/t)$" says that in *all* possible worlds *some* possible world is permitted.

"$P_3(t/t)$" says that in *some* possible world *all* possible worlds are permitted.

"$P_4(t/t)$" says that in *all* possible worlds *all* possible worlds are permitted.

The expressions of each calculus have a P_i-normal form. It is identical with the P_i-normal form of the corresponding weak calculus.

The *criterion of logical truth* in a strong dyadic deontic logic is the following one: *An expression 'E' of a strong P_i-calculus is logically true, if and only if, the P_i-normal form of the expression "$P_i(t/t) \rightarrow E$" is a tautology of propositional logic (PL).*

The strong dyadic deontic calculi too are decidable and complete relative to their truth-criteria.

In the strong P_2-calculus we prove as a theorem that $P_2(t/p)$ or that "under all circumstances something is permitted."

In the strong P_3-calculus we prove that $P_3(p/t)$ or that "under some circumstances everything is permitted."

In the strong P_4-calculus we prove that $P_4(p/q)$ or that "under all circumstances everything is permitted."

It may be thought that, because of these theorems, the *strong* P_3- and P_4-calculi are absurd. But they are not "illogical."

The strong dyadic P_1-calculus is the same as the dyadic system presented in Section I.

---

[9] If the expression of PL to the left or to the right of '/' happens to be a contradiction, it has no disjunctive normal form. In that case, replace the expression by '$\sim t$' and add '$\sim t$' as a disjunct to all the other disjunctive normal forms.

[10] See "Deontic Logic," *Mind*, vol. 60 (1951), p. 9.

VI.

The dyadic calculi which we have so far been considering accept the "interdefinability"-principle that $P_i(p/q) \leftrightarrow \sim O_i(\sim p/q)$. This principle, however, may become challenged.

We split up the principle in two implications, viz., the principle that $P_i(p/q) \rightarrow \sim O_i(\sim p/q)$ and the principle that $\sim O_i(\sim p/q) \rightarrow P_i(p/q)$. The first says that if, given that q, it is permitted that p then it is not the case that, in the same circumstances, it is obligatory that $\sim p$. This seems extremely plausible. For, is not to say that it is obligatory that $\sim p$ the same as saying that it is forbidden that p? If this is so, then to deny the first implication would be tantamount to admitting that, under given circumstances, one and the same thing could be both permitted and forbidden. But this seems contrary to the very *meaning* of "permitted" and "forbidden."

The second implication says that if, given that q, it *is not* obligatory that $\sim p$, then it *is* permitted that p. This is much less obvious than the first principle. For, it seems to say that what is not forbidden is *ipso facto* permitted. But can this be regarded as true by virtue of the *meaning* of "permitted" and "forbidden" alone?

The idea that what is not forbidden is *ipso facto* permitted is related to the principle of legal philosophy "*nullum crimen sine lege*." I shall, for convenience, refer to the principle that $\sim O_i(\sim p/q) \rightarrow P_i(p/q)$ as the *nullum crimen* idea. If we take the view that the *nullum crimen* principle does not embody a logical truth concerning the notions of permission, prohibition, and obligation, we must also refuse to regard the "interdefinability"-principle as logically true.

Deontic logics which reject A1 wholesale would hardly be of interest. But systems which accept the principle that $P_i(p/q) \rightarrow \sim O_i(\sim p/q)$ but reject the *nullum crimen* principle are certain to be of interest.[11]

If the *nullum crimen* principle is rejected, we can no longer in the calculi derive distribution laws for the notion of obligation from those for the notion of permission. The distribution laws for obligation have to be derived from the definitions of the various obligation-concepts and stated as axioms in the calculi.

Deontic logics which do not assume that the deontic operators are interdefinable, I shall call P_iO_i-calculi. The P_iO_i-calculi which correspond to the *weak* P_i-calculi have the following axioms:

A0 DP_1O_1. As before.
A1 DP_1O_1. $P_i(p/q) \rightarrow \sim O_i(\sim p/q)$. $(1 \leqslant i \leqslant 4.)$
A2a DP_1O_1. Same as A2 DP_1.
A2b DP_1O_1. $O_1(p\&q/r \vee s) \leftrightarrow O_1(p/r)\ \&\ O_1(p/s)\ \&\ O_1(q/r)\ \&\ O_1(q/s)$.
A2b DP_2O_2. $O_2(p\&q/r \vee s) \leftrightarrow (O_2(p/r)\ \&\ O_2(q/r)) \vee (O_2(p/s)\ \&\ O_2(q/s))$.
A2b DP_3O_3. $O_3(p\&q/r \vee s) \leftrightarrow (O_3(p/r) \vee O_3(q/r))\ \&\ (O_3(p/s) \vee O_3(q/s))$.
A2b DP_4O_4. $O_4(p\&q/r \vee s) \leftrightarrow O_4(p/r) \vee O_4(p/s) \vee O_4(q/r) \vee O_4(q/s)$.

The principles of inference are R1–R3.

The expressions of each calculus have what I propose to call a P_iO_i-normal form. It is obtained in the following way: In the constituent P_i-expressions we replace the expressions to the left and to the right of '/' by their disjunctive normal forms in terms of *all* the variables of the whole expression. In the constituent O_i-expressions we replace the expressions to the left of '/' by their conjunctive and the expressions to the right of '/' by their disjunctive normal forms in terms of *all* the variables of the whole expression.[12] Thereupon we distribute the constituents according to the (appropriate) axioms A2a and A2b. The result is the P_iO_i-normal form.

The criterion of logical truth in the weak P_iO_i-calculi is: An expression 'E' is logically true, if and only if, the P_iO_i-normal form of the expression

$$\text{``}\prod_{\mu=1}^{2^n} \prod_{\nu=1}^{2^n} (P_i(c\mu/c_\nu) \rightarrow \sim O_i(d\mu/c_\nu)) \rightarrow E\text{''}$$

is a tautology of PL. c_μ is a state-description in terms of the variables of E. d_μ is the negation of c_μ. $\prod_1^k$ in front of an expression of a certain form means the conjunction of all expressions of this form which we obtain by successively replacing an index-numeral by numerals 1, . . ., k.

The *strong* P_iO_i-calculi have an additional axiom:

A3 DP_1O_1. Same as A3 DP_1.

The P_iO_i-normal forms are the same as in the weak calculi.

The criteria of logical truth for the strong calculi are obtained from those of the weak calculi by adding as a conjunct "$P_i(t/t)$" to the antecedent of the implication.

---

[11] One such system is explored in some detail in my book *Norm and Action* (London, Routledge & Kegan Paul, 1963).

[12] If the expression in question has no conjunctive normal form, write in its place 't' and add 't' as a conjunct to all the other conjunctive normal forms. If again the expression has no disjunctive normal form, write in its place '$\sim t$' and add '$\sim t$' as a disjunct to all the other disjunctive normal forms.

VII

We next define *monadic counterparts* for the dyadic calculi.

The monadic expression "P_1p" will mean the same as the dyadic expression "$P_1(p/t)$" and hence, by the definition of the latter, it will mean that in *some* possible world *some* possible world is permitted in which it is true that p.

"P_2p" will mean the same as "$P_2(p/t)$" and hence it will mean that in *all* possible worlds *some* possible world is permitted in which it is true that p.

"P_3p" will mean the same as "$P_3(p/t)$" and hence it will mean that in *some* possible world *all* possible worlds are permitted in which it is true that p.

"P_4p" will mean the same as "$P_4(p/t)$" and hence it will mean that in *all* possible worlds *all* possible worlds are permitted in which it is true that p.

The axioms of the *weak* monadic P_i-calculi are:

A0 MP_i. A set of axioms of PL (with well-formed formulae substituted for the variables).
A1 MP_0. $P_ip \leftrightarrow \sim O_i \sim p$. $(1 \leqslant i \leqslant 4.)$
A2 MP_1. $P_1(p \vee q) \leftrightarrow P_1p \vee P_1q$.
A2 MP_2. $P_2(p \vee q) \leftrightarrow P_2p \vee P_2q$.
A2 MP_3. $P_3(p \vee q) \leftrightarrow P_3p \;\&\; P_3q$.
A2 MP_4. $P_4(p \vee q) \leftrightarrow P_4p \;\&\; P_4q$.

The rules of inference are R1–R3.

A2 MP_1 and A2 MP_2 have the same structure. But they have different meanings. Their difference in meaning, however, reveals itself only in their dyadic translation. Hence we could say that the P_1- and P_2-calculi are "monadically identical." The same is true of the P_3- and P_4-calculi.

The axioms of the *strong* monadic P_i-calculi are the axioms of the weak calculi with the following addition:

A3 MP_i. P_it. $(1 \leqslant i \leqslant 4.)$

The strong monadic P_1-calculus is identical with the monadic deontic logic which was described in Section I.

A monadic calculus too may reject the *nullum crimen* principle. We then get monadic P_iO_i-calculi.

The axioms of the *weak* monadic P_iO_i-calculi are:

A0 $\mathrm{MP}_i\mathrm{O}_i$. Same as A0 MP_i.
A1 $\mathrm{MP}_i\mathrm{O}_i$. $P_1p \rightarrow \sim O_i \sim p$. $(1 \leqslant i \leqslant 4.)$
A2a $\mathrm{MP}_i\mathrm{O}_i$. Same as A2 MP_i.
A2b $\mathrm{MP}_1\mathrm{O}_1$. $O_1(p \& q) \leftrightarrow O_1p \;\&\; O_1q$.
A2b $\mathrm{MP}_2\mathrm{O}_2$. $O_2(p \& q) \leftrightarrow O_2p \;\&\; O_2q$.
A2b $\mathrm{MP}_3\mathrm{O}_3$. $O_3(p \& q) \leftrightarrow O_3p \vee O_3q$.
A2b $\mathrm{MP}_4\mathrm{O}_4$. $O_4(p \& q) \leftrightarrow O_4p \vee O_4q$.

The axioms of the *strong* monadic P_iO_i-calculi, finally, are those of the weak calculi with the addition:

A3 $\mathrm{MP}_i\mathrm{O}_i$. Same as A3 MP_i.

Again the monadic P_1O_1-calculus is "monadically identical" with the monadic P_2O_2-calculus, and the P_3O_3-calculus with the P_4O_4-calculus.

Every expression of the weak and strong monadic P_i-calculi has a P_i-normal form and every expression of the weak and strong monadic P_iO_i-calculi has a P_iO_i-normal form. The normal forms are obtained by replacing expressions of PL by their disjunctive normal forms in P_i-expressions and conjunctive normal forms in O_i-expressions and distributing the expressions. (In the P_i-calculi we first replace O_i-expressions by the equivalent P_i-expressions.)

The criteria of logical truth of an expression 'E' are:

In the weak monadic P_i-calculi: that the P_i-normal form of 'E' be a tautology of PL.

In the strong monadic P_i-calculi: that the P_i-normal form of "$P_it \rightarrow E$" be a tautology of PL.

In the weak monadic P_iO_i-calculi: that the P_iO_i-normal form of

$$``\prod_{\mu=1}^{2^n} (P_{ic\mu} \rightarrow \sim O_{id\mu}) \rightarrow E"$$

be a tautology of PL.

In the strong monadic P_iO_i-calculi: that the P_iO_i-normal form of

$$``P_it \;\&\; \prod_{\mu=1}^{2^n} (P_{ic\mu} \rightarrow \sim O_{id\mu}) \rightarrow E"$$

be a tautology of PL.

VIII

The four P-concepts which we distinguished in Section IV are not logically unrelated. As seen from their definitions, $P_4(p/q)$ entails $P_3(p/q)$ and $P_2(p/q)$ and $P_1(p/q)$; $P_3(p/q)$ and $P_2(p/q)$ both entail $P_1(p/q)$; $P_3(p/q)$ and $P_2(p/q)$ are independent of one another. These relations carry over to the monadic concepts.

To the four P-concepts there correspond four O-concepts. $O_1(p/q)$ entails $O_2(p/q)$ and $O_3(p/q)$ and $O_4(p/q)$; $O_2(p/q)$ and $O_3(p/q)$ both entail $O_4(p/q)$; $O_2(p/q)$ and $O_3(p/q)$ are independent.

It is of some interest to note that to the *weakest* concept of permission, there answers the *strongest* concept of obligation—and to the strongest obligation the weakest permission.

There is no objection to forming "mixed" truth-functional compounds of P_i- and/or O_i-expressions with different indices i—nor to forming compounds of dyadic and monadic expressions. When contemplating the logical features of such "mixed" expressions we must be clear over which axioms we assume for the various P- and O-concepts, viz., whether we accept or reject the Principle of Permission and whether we accept or reject the *nullum crimen* principle for them. We may accept these principles for some of the P- and/or O-concepts involved and reject them for others.

The P_i- and/or P_iO_i-normal forms of the "mixed" expressions we get by transforming each one of the constituent expressions into *its* P_i- or P_iO_i-normal form.

In formulating criteria of truth for a calculus containing "mixed" expressions we have to observe the following equivalences which reflect the dependence between the various P- and O-concepts:

(1) $P_4(p/q) \rightarrow P_3(p/q)$
(2) $P_4(p/q) \rightarrow P_2(p/q)$
(3) $P_4(p/q) \rightarrow P_1(p/q)$
(4) $P_3(p/q) \rightarrow P_1(p/q)$
(5) $P_2(p/q) \rightarrow P_1(p/q)$
(6) $O_1(p/q) \rightarrow O_2(p/q)$
(7) $O_1(p/q) \rightarrow O_3(p/q)$
(8) $O_1(p/q) \rightarrow O_4(p/q)$
(9) $O_2(p/q) \rightarrow O_4(p/q)$
(10) $O_3(p/q) \rightarrow O_4(p/q)$

IX

We can now deal afresh with the "paradoxes" of deontic logic which were mentioned in Section II. They all arise, we said, in the monadic system presented in Section I. This system is identical with the *strong* monadic P_1-calculus. It is easily shown, however, that the very same paradoxes arise already in the *weak* monadic P_1-calculus and that they occur in the weak and strong P_1O_1-calculi too. For their derivations depend only upon the distributive properties of the concepts P_1 and O_1 and are independent of acceptance of the Principle of Permission and the *nullum crimen* principle.

It is obvious that none of the paradoxes which we mentioned arise in the monadic P_3- and P_3O_3-calculi.

Ross's paradox does not arise, since O_3p does *not* entail $O_3(p \vee q)$ in any of the calculi. Thus, in particular, if to mail a letter is O_3-obligatory, it does not follow that to mail this letter *or* to burn it is O_3-obligatory.

Since $\sim P_3p$ does not entail $O_3(p \rightarrow q)$ nor $\sim P_3(p \& q)$, the paradoxes of derived obligation and of "the good Samaritan" do not arise in the P_3- or in the P_3O_3-calculi.

In the monadic P_3- and P_3O_3-calculi the equivalence $P_3(p \vee q) \leftrightarrow P_3p \& P_3q$ holds. This is not here connected with any paradox, not to speak of contradiction.

Shall we, on the basis of these findings, conclude that the P_3- and P_3O_3-calculi are those which best agree with our intuitive ideas of permission and obligation? The answer is that we must *not* conclude this. For, in these calculi too "paradoxes" arise. Thus we can in them prove, e.g., the formula $P_3p \rightarrow P_3(p \& q)$. It says that if it is permitted that a state of affairs obtains, then it is also permitted that any conjunction of this state with another obtains. Thus, in particular, a permission to enjoy life would entail a permission to enjoy life *and* commit murder (though not a *permission* to commit murder). This paradox may be said to correspond in the P_3-calculi to Ross's Paradox in the P_1-calculi. One can meet the paradox with similar arguments to those which have traditionally been used to meet Ross's Paradox in the standard systems of deontic logic. (See above Section II.) But one can also refuse to let those arguments be the end of the story and look for a calculus in which the paradoxical formula simply cannot be proved. Such a calculus is the (weak and strong) P_1-calculus.

Thus, in a way, the answers to the paradoxes which arise in the P_1-calculi are given in the P_3-calculi—and the answers to the paradoxes which arise in the P_3-calculi are given in the P_1-calculi.

X

The moral to be drawn from all this is that there are *several concepts* of permission and obligation and that the "paradoxes" of deontic logic arise from a confusion of them. When the concepts have become clearly distinguished and systematized, there are no "paradoxes."

To this insight should be added an observation concerning ordinary use of language. It seems to be the case that "or" in connection with the phrase "permitted that p or q" is normally so used that it suggests a P_3-permission, viz., a permission the sense of which is that anything which falls under (entails) the alternative "p or q" is permitted. And it also seems to be the case that "and" in connection with the phrase "obligatory that p and q" is normally so used that it suggests an O_1-obligation, viz., an obligation the meaning of which is that anything under which the compound "p and q" falls, i.e., anything entailed by it, is thereby also obligatory. Thus conjunctive obligations and disjunctive permissions are "naturally" understood in the strongest sense of "obligation" and "permission." But since the correlative in a deontic logic to the strongest obligation-concept is the weakest permission-concept, it takes different "deontic logics" to do justice to all our intuitions in the matter.

Appeal to "intuition" in discussions of the acceptability or validity of various formulae of monadic or dyadic deontic logic is on the whole futile. When the question of validity is at stake, it should be framed as follows: In which deontic logic, if in any, is this formula valid? To each deontic logic answers different concepts of permission and obligation. The question can therefore also be formulated: For which concept of permission and/or obligation is this formula valid? We can trivially construct permission- and obligation-concepts, and therewith deontic logics, so that all formulae except those which can become refuted in some non-deontic branch of logic, become valid. But not all such constructions are interesting and worthwhile.

The Academy of Finland

Received April 15, 1966

American Philosophical Quarterly
Volume 4, Number 2, April 1967

VI. JUSTICE REASSESSED

D. C. EMMONS

THE concept of justice has received a great deal of attention from both classical and contemporary moral theorists. In what follows I shall sketch a logical map of the root notion of justice and show that many ethicists distort this notion, either by relating it improperly to other important moral concepts or by offering specious arguments for the obligation to be just. I shall argue that the injunction to be just is a logically independent moral demand on a par with a wide range of competitors, despite the fact that justice is in a sense parasitic on benevolence (or maleficence).

I

The generic concept of justice can be analyzed as follows:

> An action (or class of actions, or institution) is just if (and only if) it results in an equal distribution of goods and/or evils to equals.

Three points must be made at the outset. First, this definition is deliberately abstract, covering all forms of distributive justice. It may not correspond to all those senses of "justice" in common use, but it is I think the most basic.

Secondly, the definition I have proposed must be supplemented by a criterion for equality before it can be applied; and it permits us to use any of a *plurality* of concrete criteria. It might be argued, by those who reject all forms of retributive justice, that it is improper to parcel out rewards or penalties in accord with the past activities of persons or institutions. Such a claim is often phrased: "That would not be *real* justice." This strikes me as a verbal error, and parallels the error made by those who say of the profligate's life that it does not constitute *real* happiness. Both of these semantic revisions obscure the fact that the dispute is normative rather than verbal. It is much less confusing to say that the profligate's life is not the best he could realize, even though he may be happy. And to say that retributive justice is morally wrong, even though it is a form of justice. In sum, I feel it is advisable to utilize a generic definition of "justice" which covers a number of substantive criteria, as determined by various criteria for equality or merit. The definition suggested above prevents us from begging genuine moral questions.

Thirdly, it is useful to note that justice is in a special sense parasitic on the notion of benevolence (or maleficence). The obligation to be just by no means follows logically from the obligation to be benevolent. Quite the contrary, as I shall argue below. But there is a connection. Justice concerns the distribution of goods and evils, while beneficence (and maleficence) involve the doing of good (and evil). So beneficence (and maleficence) are necessary, but not sufficient, conditions for justice. Alternatively, it must be possible to benefit or harm people before questions of justice can arise. This may explain why some classical moral theorists (erroneously) thought justice to "be included in" the notion of benevolence or utility.

I wish to refine the concept of justice further by defining two more specialized concepts falling under it:

(1) *Perfect justice*: a distribution of goods and evils in direct proportion to merit, where each receives a share.

(2) *Imperfect justice*: a distribution of limited and/or non-divisible goods and evils as follows: (a) Since all cannot have a share, some are neglected. But those who are so deprived are chosen by lot. (b) The randomness of deprivation is extended over time. If, e.g., a unit good is awarded to an individual after the first lottery, he is excluded from further lotteries until every member of the population has received such a good. Then the process repeats.

It should be obvious that both perfect and imperfect justice fall under the generic definition. The difference is a function of environmental factors only, and the inequalities of imperfect justice are as "fair" as circumstances permit.

I have distinguished perfect and imperfect justice to make it clear that conditions over which we have no control need not rule out justice in the generic sense. Equals can always be treated as equals provided we randomize natural discrimination; hence the familiar "Let's flip a coin" and "It's my turn now."

I now wish to establish what was merely asserted above; viz., that justice is logically independent of and may conflict with other moral demands. Few ethicists have denied that justice is independent of such moral imperatives as the obligation to be kind, to revere one's elders, never to take a human life, etc. All of these could, although they need not, conflict with the obligation to be just.

But some have argued that justice cannot conflict with benevolence (i.e., the obligation to maximize over-all satisfaction). For the time being, it will be sufficient to describe one clear case of incompatibility. Assume a man who has two and only two options. He can distribute x units of satisfaction to each of n equals, with no other results, or he can award y units of satisfaction to k of the n, neglecting the rest and producing no other effects. Assume further that $xn < yk$. The first option is just, but the second is more benevolent, and the clash is irreducible. I need not belabor the point by adducing concrete examples from everyday life. Suffice it to say that moral and political indecision is often caused by precisely such clashes between the conflicting claims of justice and logically independent moral goals.

Let us turn now to discussions of justice by a series of actual and hypothetical moral philosophers. A critical survey of their views should be instructive.

II

Act-utilitarians have taken a number of positions on justice. There are three principle options open to advocates of the principle of utility, but first a word on act-utilitarianism *per se*. The act-utilitarian subscribes to a moral first principle which reads: Always perform that action which will have the best over-all results, and if there is no optimific act, perform one of the set which will have equally desirable results.

Two points must be clarified. This abstract definition of act-utilitarianism leaves open the nature of good results and their relative importance. I do not wish to grapple with this issue: it requires normative judgment rather than mere conceptual analysis. The act-utilitarian, as circumscribed above, may designate literally anything as a good or bad result, and weigh it as he sees fit. This might be construed as a fatal weakness, on either of two counts. It might be objected that a moral theory which entails a comparison of things as disparate as pleasure and knowledge is hopelessly muddled. I do not feel that this is a fatal objection, since all of us are forced to make such comparisons daily. And any number of quite legitimate moral theories entail the same kind of judgment; e.g., several variants of rule-utilitarianism and intuitionism.

It might also be argued that any moral theorist who considers things other than states of living beings when computing the net worth of an action is quite confused. How can a beautiful sunset have *intrinsic* value? Good or bad to whom? Morality was made for man and not the reverse. While this objection has persuasive force, I do not feel it is *logically* compelling. Moore saw nothing odd in ascribing intrinsic worth to inanimate states, and I would prefer to characterize the man who shares his views as a statistically abnormal moralist rather than a conceptually confused moral philosopher. To reject such an ethical stance on *a priori* logical or linguistic grounds may be to beg the question.

The second preliminary point is quite important. An act-utilitarian, as defined above, need not advocate injustice. As we shall see below, he can always take it into account when weighing intrinsic goods. While this does not eliminate the clash between most versions of act-utilitarianism and strict egalitarianism, the act-utilitarian is not biased in principle.

(A) A classical or hedonistic act-utilitarian can solve the problem of clashes between benevolence and justice by treating the latter as morally irrelevant. He claims that there is one and only one intrinsic good: happiness or satisfaction. Hence he enjoins the act which maximizes satisfaction, and this leaves room for conflicts between justice and utility. In defense of his disregard for justice, the hedonistic act-utilitarian might respond as follows: "Human happiness and misery are the only factors of moral relevance. Why act in one way rather than another if not to increase net satisfaction? I categorically deny that justice must be done though the heavens fall. Such an attitude reflects a moral superstition to the effect that there are intrinsic goods other than human welfare. I reject it."

Now such a stance would be unusual, but not incoherent. It is tempting to object that the hedonistic act-utilitarian contradicts himself by first asserting the paramount importance of human happiness and then treating justice as irrelevant. Doesn't justice concern the proper distribution of goods and evils, and hence the happiness and misery of concrete human beings? Yes and no. Justice does involve the geography of human welfare, but the hedonistic act-utilitarian posits only one goal—maximum human satisfaction. It follows that he is not guilty of self-contradiction when he refuses to sacrifice utility for justice. And if we insist upon justice when this will not detract from over-all happiness, the classical act-utilitarian can always claim that our moral injunction is irrational. On his view one unit of creature-satisfaction is the equal of any other and the question of distribution is beside the point.

At this point it is common for critics to object that the hedonistic act-utilitarian has made a simple error. Surely justice is morally relevant. How else explain the universal concern for justice as well as happiness *per se*? This move is persuasive, but it is crucial to recognize that it constitutes a substantive moral judgment rather than "a logical insight." Disputes over "relevance" in ethics are moral controversies and not mere logical or linguistic squabbles. To claim that a Negro should not be disadvantaged because of the "irrelevant" fact of skin color is to make a *moral* claim. Those who attempt to refute the hedonistic act-utilitarian in this way only obscure the real issue. The conflict between the hedonist and one who posits a multiplicity of intrinsic goods or morally relevant factors is normative. The former cannot prove that justice is ethically irrelevant by appealing to the law of non-contradiction or linguistic conventions alone. But by the same token, he cannot be discredited on such grounds. The conflict reflects a difference in ultimate moral norms.

(B) A second group of hedonistic act-utilitarians, including some influential moral theorists, tries to escape the dilemma posed by clashes between justice and their variant of utility by arguing that such conflicts are merely apparent. They assert that justice is "built into" utility (can be deduced from it). Justice *is* parasitic on utility, and this may explain why some act-utilitarians have tried to equate the two or infer the former from the latter. But this is clearly mistaken. To see more clearly why it's mistaken, it will be useful to examine attempts to establish the alleged logical connection.

One common argument is this: Justice increases net human happiness because most people have a conditioned distaste for an unjust distribution of satisfaction or dissatisfaction, and, moreover, produce more when justice obtains; hence the injunction to be just is entailed by the principle of utility plus facts about human nature.

This argument is patently unsatisfactory. It permits the hedonistic act-utilitarian to accept some demands for justice, but not all. Even if the presence (or absence) of justice had no effect whatever on net human happiness it would be legitimate to press for justice quite independent of satisfaction.

A second move is open to the act-utilitarian who wishes to deduce justice from benevolence. Assume there are two men who have equal ability to cash an extrinsic good into satisfaction. Assume further that neither can cash more than one such instrumental good into additional personal satisfaction. Under these circumstances it would be inconsistent for the hedonistic act-utilitarian to favor one over the other. Hence justice is logically entailed by utility; and thus, the two do not conflict.

This argument, while logically impeccable, also fails to eliminate all conflicts between justice and utility. Some cases of injustice do entail a loss of satisfaction; but hardly all. It is quite easy to conceive of cases such that an unjust distribution of instrumental goods would not involve such a loss due to the fact of diminishing returns. Further, this argument fails to account for the temporal element in the demand for justice as outlined above. One can often justify an equitable distribution of goods at a given time by appealing to the capacity of particular persons to cash such goods into satisfactions, but this argument cannot be used to justify the demand for equal treatment over an extended length of time.

I conclude that the attempt to deduce justice from utility is flawed in principle. In effect, the hedonistic act-utilitarian opts for a very narrow criterion for justice. He appeals to a one-dimensional criterion for equality; viz., the capacity of individuals to use a given extrinsic good for purposes of personal satisfaction. But this is hardly the only criterion open to us. It completely neglects both equality over time and logically independent standards for merit.

(C) Finally, the act-utilitarian can accommodate the demand for justice by treating it as a separate and distinct intrinsic good to be weighed against pleasure, contractual fidelity, increased knowledge, etc. Such "ideal" act-utilitarians hold that we have a *prima facie* obligation to be just which is only canceled by an opportunity to realize some other desirable result of greater weight.

Such a stance does constitute an improvement over hedonistic act-utilitarianism. It neither neglects justice altogether nor attempts to assimilate the obligation to be just to a more general duty to maximize over-all satisfaction. Further, this form of utilitarianism need not enjoin obligations which are not accompanied by correlative rights. The obligation to be just is always an obligation to some individual or group, and thus cannot be rejected out of hand by someone who refuses to recognize ideals which have no relation whatever to human welfare.

I conclude that ideal act-utilitarianism does least violence to the logic of justice and the moral convictions of ethicists and plain men alike. But there are further difficulties which should not be overlooked. As we shall see later, the act-utilitarian is open to the charge of inconsistency. Further, the ideal act-utilitarian is really an act-intuitionist in disguise, but lacks some of the latter's virtues. To see this, we must turn to the intuitionist's treatment of justice.

III

I shall attempt no more than a cursory summary of intuitionism *per se*. Some intuitionists argue that we simply "recognize" the rightness of actions having certain empirical properties and then generalize the results of these isolated moral intuitions. Moral rules are treated as rules of thumb —mere summaries of past judgments of individual actions. Others suggest that we "recognize" the truth of injunctions of the form "actions of type X are *prima facie* right (or wrong)." This difference is not important in practice, although rule-intuitionism cannot degenerate into the bizarre claim that each and every act is unique and must therefore be judged independent of its "gross" similarities to other acts. Such a view distorts the significance of a trivial logical truth; viz., particular actions must have individuating characteristics. But such characteristics may reduce to spatio-temporal location, and few are prepared to assert that mere differences of time or place are morally relevant. It is not of course self-contradictory to argue in this way, but it hardly corresponds to the facts of everyday moral deliberation. In fact, it rules out genuine moral deliberation, for it prohibits judgments of the form "action X is right (or wrong) *because* it has features a, b, . . ., n." If such an argument is necessarily incomplete because all actions are unique, it is futile to support a moral judgment by appealing to moral rules or factual reasons. Surely this is an extraordinary claim.

Suffice it to say that almost all intuitionists agree on a number of central points. They assert that we have a number of distinct and potentially conflicting *prima facie* obligations. Among them is the obligation to be just. They go on to argue that we must simply weigh such duties when they clash, and perform the action which is more obligatory than any alternative open to us. Some intuitionists acknowledge that it would be convenient to have second-order rules which rank our obligations, but they doubt that such rules can be discovered.

This brief sketch of intuitionism will serve to explain why I described the ideal act-utilitarian as an act-intuitionist in disguise. I must now explain why I prefer the latter over the former; both assert the obligation to be just and recognize that it is logically independent of other widely accepted obligations. But the ideal act-utilitarian obscures the full difference between duties of benevolence and duties of justice. He describes the situation in terms of an overriding obligation to maximize the good, where "the good" is made up of a number of independent parts. This is somewhat misleading, and conflicts with the ordinary language connotations of the phrase "doing good." As most of us use this locution it suggests bestowing benefits on someone; and nothing else. But justice often involves benefiting A at the expense of B. Hence it is not so much "doing good" as "doing good in the right way."

So much for the virtues of the intuitionist's treatment of justice. Are there any weaknesses? There is no need to repeat the criticisms that have been leveled at intuitionism on general epistemological grounds,[1] nor to belabor the fact that intuitionists seldom depart widely from "common sense" moral judgments. There is another apparent flaw more germane to our purposes. I have praised the intuitionist for his refusal to blur important distinctions in the quest for simplicity and/or system. It is arguable, however, that he overlooks the possibility that our alleged obligations reduce to two or three generic duties. Intuitionists rarely posit obligations which have nothing to do with tangible human goods or their proper distribution. Hence it seems reasonable to suggest that the non-naturalist, *qua* normative ethicist, recognizes only two classes of duties; namely, those of benevolence and those of justice. Why distinguish separate duties of fidelity, reparation, gratitude, and self-improvement[2] if these can be assimilated to beneficience or justice (or rejected because there are no genuine duties to oneself)? I suggest that we attempt to steer a middle course between the misleading conflations of the hedonistic act-utilitarian and the equally misleading proliferation of intrinsic goods or *prima facie* obligations recognized by the ideal act-utilitarian or the intuitionist.

---

[1] Cf. P. F. Strawson, "Ethical Intuitionism," *Philosophy*, vol. 24 (1949), pp. 23–33 and P. H. Nowell-Smith, *Ethics* (London, Penguin Books, Ltd., 1954), ch. 3 and 4.

IV

Rule-utilitarians (or universalists) attempt to remedy a serious defect in act-utilitarianism. It is arguable that they succeed. It is not at all obvious that the improvement constitutes a recognition of the demands of justice. To see this we must first clarify the nature of rule-utilitarianism and contrast it with act-utilitarianism. According to the latter, the crucial question for a moral agent takes the form: "What will happen if I do *X*, or *Y*, or *Z*?" The universalist contends that the relevant question is quite different; viz., "What will happen if everyone like me does *X*, or *Y*, or *Z* in similar circumstances?"

Now which of the two is correct, and why? The rule-utilitarian asserts that act-utilitarianism is internally inconsistent. His argument runs as follows: "Act-utilitarians wish to achieve maximum over-all good. They enjoin every agent to perform *the* action which will produce the best result. But this leads to an inconsistency whenever: (a) a particular course of action is optimific for each member of a class, but (b) the collective consequences of such behavior are less than optimal or disastrous. Hence the appropriate query for a moral agent must be: 'What will happen if everyone . . .?' This question always yields a rational answer, whereas the act-utilitarian's simple test often yields a grotesque conclusion."

Examples will help to clarify the rule-utilitarian's position. It would often be optimific for an agent to lie, break a promise, steal, fail to obey the law, etc. But if everyone like the agent were to do so, the results would be undesirable. Hence simple act-utilitarianism is unsatisfactory. We must investigate the consequences of collective patterns of behavior rather than isolated acts. Morality is a cooperative enterprise, and self-defeating moral injunctions result from a failure to recognize this crucial fact.

I feel that the rule-utilitarian's critique of act-utilitarianism is basically sound. This conclusion is strengthened by the fact that act-utilitarians often depart from the principle of utility when faced by cases like the ones noted above[3] or offer specious arguments to show that the results of a particular white lie or theft would not really be optimific. Suffice it to say that the act-utilitarian contradicts himself whenever he recommends the optimific act even though the collective consequences of such acts would be undesirable. The more interesting question is this: Does rule-utilitarianism entail a demand for justice neglected by the act-utilitarian?

It might appear as if the answer is affirmative. Isn't it "unjust" to steal or lie, even though the results in a particular case might be best? Isn't it "unfair" for me to cheat on my income tax, despite the fact that this will benefit me and mine and will not really undermine the economy? What if everyone did such things? This analysis breaks down on closer inspection. Act-utilitarianism is not inherently unjust, nor is rule-utilitarianism necessarily equitable. Consider the alleged obligation to pay one's taxes. The consistent act-utilitarian must enjoin that we disregard this widely accepted duty whenever it would be optimific to do so. Now this may be self-defeating, but it is not to recommend injustice in the sense defined earlier. What if the money so obtained were used to benefit all of mankind with complete impartiality? As we saw above, the act-utilitarian is not an egoist. In fact, he need not subscribe to any kind of systematic bias. The hedonistic act-utilitarian does overlook the demand for justice *per se*, but the ideal act-utilitarian need not.

Further, it is a confusion to assert that the rule-utilitarian necessarily embeds the demand for justice in his normative ethical code. It all depends upon his criteria for "good" and "bad" results. To put it tersely, the hedonistic rule-utilitarian is no more committed to the goal of justice than the hedonistic act-utilitarian. Now the ideal rule-utilitarian may enjoin justice as a *prima facie* moral obligation, but if so this is a consequence of his concrete prescriptions rather than the formal characteristics of his system.

---

[2] Cf. W. D. Ross, *The Right and the Good* (Oxford, The Clarendon Press, 1930).

[3] There is no better exemplar than Mill, who asserts in a moment of insight and/or weakness: "In the case of abstinences, indeed—of things which people forbear to do from moral considerations, though the consequences in the particular case might be beneficial—it would be unworthy of an intelligent agent not to be consciously aware that the action is of a class which, if practiced generally, would be generally injurious, and that this is the ground of the obligation to abstain from it" (*Utilitarianism* [New York, The Liberal Arts Press, Inc., 1957], p. 25). Quite so, but this is a blatant departure from orthodox act-utilitarianism.

This result should not be surprising. If rule-utilitarianism is superior to act-utilitarianism on wholly *a priori* grounds, and if justice is a substantive moral demand independent of benevolence *et al*, the rule-utilitarian can hardly claim that he necessarily enjoins a concrete obligation which his naive cousin neglects. As the novice in logic soon learns, mere consistency does not guarantee material truth. In more relevant terms, logical rigor alone does not protect us against moral oversight. The universalist may, but need not, subscribe to the dictum: "Let justice be done though the heavens fall." And the act-utilitarian may, but need not, subscribe to the injunction: "Maximize human happiness and justice be damned." Finally, it is hardly obvious that the latter is inferior to the former. Extremism in the defense of justice is not a self-evident virtue.

V

I wish to turn now to the arguments of those who appeal to logic or the meanings of terms to show that we have an obligation to be just. I shall attempt to show that all such arguments constitute either an illegitimate use of "persuasive definitions" or a question-begging inference from "is" to "ought."

(A) Some philosophers claim that morality involves "doing things on principle." They go on to assert that consistent moral principles entail justice, arguing that self-contradictory maxims necessarily embody *ad hoc* exceptions in favor of some individual or group. Our discussion of rule-utilitarianism should suggest both the appeal and the flaw in this line of argument. Morality does involve doing things on principle. And it is arguable that inconsistent moral principles are *a priori* unsatisfactory.[4] It is also true that inconsistencies in a normative ethical system often take the form of biased exceptions for or against some agent or agents; but this need not be the case. Such inconsistencies often constitute unjust discrimination, but they may also enjoin action which would fail to produce maximum over-all satisfaction, fidelity to the *n*th commandment, or none of these. In brief, simple inconsistency does not entail injustice. And quite consistent moral principles like the injunction to maximize happiness or secure a minimal level of welfare for all may entail injustice. In sum, neither consistent nor consistent and universalizable moral principles guarantee substantive justice.

(B) Other moral philosophers argue that the notion of justice is built into the meanings of "ought" and "morality." They assert that justice is a necessary condition for right action, and support this claim by referring to the generalization principle.[5] The GP reads as follows: What is right (or wrong) for one agent is right (or wrong) for all like agents in like circumstances.

I suggest that the GP is built into the meanings of "ought" and "morality" and hence constitutes a necessary condition for right action if and only if it is couched in terms of "reasons." Then it reads: The reasons which show that an action is right for one agent necessarily entail that such an action is right for all like agents in like circumstances. However, this form of the generalization principle does not constitute an implicit injunction to be just. Assume a group of *n* individuals in identical circumstances. Assume further that they can achieve maximum over-all happiness, at the cost of justice, if and only if each performs action *X*. Now the reasons which dictate *X* for one such agent dictate it for all. In effect, the goal of maximum over-all satisfaction dictates *X* for each because it dictates *X* for all. But only at the cost of justice, which need not be taken into account. So the generalization principle does not necessarily enjoin like treatment of like agents. If couched in terms of reason, it dictates nothing more than consistent moral recommendations; i.e., like *behavior* for all like agents in like circumstances. And this need not produce either perfect or imperfect justice.

Advocates of the GP might protest that I have distorted or emasculated this moral principle. They might assert that it is built into the meanings of "ought" and "morality" in the following form: If it is right to treat an agent in a given way it is right to treat all like agents in the same way in like circumstances. Now *this* moral principle does indeed constitute a demand for justice. But I deny that it is necessary and/or part of the meanings of "ought" and "morality." It cannot be denied that egalitarianism has been rejected by many. Often this has been done without any justification, but it does not follow that such systems do not qualify as genuine moral theories. Further, the independence of benevolence and justice permits one to formulate a consistent moral theory in which the demand for injustice is justified. In brief, there is nothing to prevent a normative ethicist from asserting that the claims of justice should be superceded by some other moral demand.

---

[4] Although it is not so obvious that this or any other formal defect necessarily entails immorality. Cf. below.

[5] See Marcus Singer, *Generalization in Ethics* (New York, Alfred A. Knopf, 1961), chaps. II and III.

I claimed above that the injunction to be just has not been embedded in the very meanings of "ought" and "morality." This might be disputed as follows: "Many crucial normative terms do entail substantive prescriptions. Take the term 'happiness.' In ordinary language 'happiness' is synonymous with 'the good life.' Hence it's analytic that we ought to pursue happiness. But it's also analytic that no man is happy if he's suffering intense and continuous pain. Hence it's analytic that we ought to avoid such a condition. Analogously, the terms 'ought' and 'morality' have concrete implications for action. It's trivially true that one ought to be moral. And one ought, logically, to treat like creatures alike. So the demand for justice is built into the very meaning of 'ought' (the very concept of 'morality')."

This appeal to language is patently specious. It may be true that we have embedded a substantive value judgment in the meaning of the word "happiness." This hardly sanctions the statement: "It's silly to deny that we ought to avoid intense and uninterrupted pain." Such a denial would be statistically odd and linguistically deviant, but hardly self-contradictory. Now I doubt that we have packed a genuine moral demand into the meaning of "ought" or "morality," but even if we have this proves nothing. Such terms, where they exist, constitute "persuasive definitions" in Stevenson's sense. They embody widely-held normative attitudes in addition to mere linguistic conventions. And it is obvious that they obscure things by making it appear as if the moral reformer is simply quibbling over words. This is unfortunate, and it would be better to eliminate all terms with both normative and descriptive content. They constitute a particularly subtle form of the naturalistic fallacy.

(C) I have tried to show that no appeal to logic and/or language demonstrates conclusively that we have an absolute obligation to be just. A weaker claim, and its supporting argument, might run as follows: "I concede that the obligation to be just is a substantive duty which might conflict with other duties. Hence it can't be absolute. But we do have a *prima facie* obligation to be just. Justice involves an equitable distribution of benefits and liabilities. Hence justice has the support of reason. But injustice, all other things equal, is inherently unreasonable, because it involves benevolence to one person at the expense of another for no reason at all. But injury to or deprivation of someone without justification is *a priori* immoral. Hence injustice is *prima facie* wrong." This argument is persuasive, but not logically conclusive. It is always open to a dissenter to deny the alleged *prima facie* obligation on the grounds that the contrast between justice and injustice is a difference that doesn't make a difference. The moralist who argues in this way displays a minority attitude, but not incoherence. He simply refuses to acknowledge that there are ethically compelling reasons for justice *per se*. And to accuse him of "moral blindness" is to beg the question by simply reiterating a concrete ethical judgment.

Summarizing, no logically binding moral argument can be built on linguistic or logical premisses alone. Facts of language do not entail moral conclusions because what we say does not dictate what we ought to do. Analogously, the fact that a moral agent reasoned inconsistently, or acted inconsistently over a period of time, or acted for no reason at all does not imply that he did something morally wrong. Now it could be argued that the man who formulates an internally inconsistent moral system is guilty of a philosophical lapse because such a system entails conflicting imperatives. But it does not follow that he is guilty of a moral lapse. To show this one would have to invoke a substantive moral premiss of the form: "Such 'reasoning' is blameworthy because it leads to immoral behavior and/or influences others to behave immorally." And to justify this one would have to appeal to a higher order moral principle together with factual premisses about the behavioral effects of muddled thinking. Further, it is always open to another ethicist to *recommend* confused moral deliberation (or the absence of moral deliberation), on the grounds that such "reasoning" results in spontaneous or unpredictable behavior and hence produces the best of all possible worlds.

It follows that neither an absolute nor a *prima facie* obligation to be just can be deduced from logical or linguistic considerations alone. Since the demand for justice is a concrete moral demand it can only be supported by describing the kind of world that would result if men were just, and recommending such a world as best. It is confused or disingenuous to argue that pure reason compels the man of good will to pursue justice. This is a highly misleading way of making a moral claim that does not differ in principle from the injunction to alleviate suffering, promote pleasure, or wipe Communism from the face of the earth.

VI

Our discussion of justice (i.e., the equitable treatment of equals) has established the following points:

(A) The moral injunction to be just is logically independent of, and hence on a par with, any number of potentially conflicting moral demands. It is, however, parasitic on benevolence (and maleficence) in the sense that justice presupposes the possibility of doing good or evil to persons.

(B) Act-utilitarians either neglect justice altogether, attempt to deduce it from benevolence, or treat it as a separate intrinsic good. Anyone who categorically denies the moral relevance of justice is statistically abnormal; but not conceptually incoherent. Those act-utilitarians who claim that justice can be inferred from benevolence are mistaken; it cannot (logically) be done. Ideal act-utilitarians usually recognize the obligation to be just and treat it as a duty quite distinct from benevolence. But they obscure the fact that justice is best described as a distributive moral good.

(C) Intuitionists, in many ways the least muddled on the logical status of justice, both assert a *prima facie* obligation to be just and refuse to assimilate it to other alleged moral goals. Further, they maintain the useful distinction between duties of benevolence and duties of justice. It is arguable, however, that intuitionists overlook possible simplifications in their zeal to formulate a normative ethical system which accurately represents the diversity of morally relevant facts. Finally, act-intuitionists deny or blur the importance of the question "What if everyone . . .?" Like the act-utilitarians, they distort the nature of moral rules. Some may be mere rules of thumb based upon particular "moral intuitions," but others embody our convictions about collective action of one sort or another.

(D) Rule-utilitarians correct a concealed but genuine inconsistency in orthodox act-utilitarianism. This formal improvement does not, however, constitute a substantive demand for justice, as some rule-utilitarians suggest. Mere consistency does not entail any moral injunction. Hence it does not entail justice.

(E) Those normative ethicists who attempt to justify the demand for equitable treatment by appealing to language or logic seriously distort moral judgments of the form: "One ought to be just." This ethical injunction neither reduces to, nor follows from, any logical or linguistic imperative. Hence it is specious to assert that concrete forms of injustice like slavery are *a priori* wrong because they are inconsistent with the essence of morality or cannot be prescribed without contradiction. It is both more accurate and more effective to argue against the advocate of discrimination on moral grounds, because he can refute or trivialize charges of logical error or verbal unorthodoxy. Injustice is no more contrary to reason than pain or death, and those who oppose it would do well to recognize this fact. Philosophical sleight of hand only weakens an otherwise defensible moral posture.

C.W. Post College of
Long Island University

Received February 28, 1966

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 2, April 1967

VII. THE PROBABILITY OF THE SIMPLE HYPOTHESIS

GEORGE SCHLESINGER

I

THE current trend among philosophers of science is to treat with suspicion any attempt to justify ontologically the principle of simplicity. Apart from the negative reason that there is no basis for assuming that nature is governed by simple laws, there are two positive arguments to support this attitude: one empirical, the other logical.

First, it is claimed that in the light of the findings of modern science and in the face of its ever-increasing complexity, it is absurd to maintain that nature itself prefers simple laws. We no longer believe, for instance, that the co-variation of the pressure and the volume of a given gas is truly represented by Boyle's simple function, and we know that Kepler's simple laws of planetary motion are rough approximations only, as in fact planets do not move along elliptic orbits but along extremely complicated curves.

The second logical reason is this: Given a number of parameters there is only limited freedom in arbitrarily fixing the relationships among them. Once we have stipulated what form a part of those relationships should take, the form of the rest inevitably imposes itself upon us. Thus if we stipulated that planetary interactions are governed by the simple laws of Newtonian Dynamics, then it inevitably follows that where more than two celestial bodies interact the resulting orbit cannot be a simple elliptic orbit for any one of them. Again, if we stipulate that the simple laws of the Kinetic Theory govern the interplay of molecules, then if the molecules have a finite volume it is easily shown that the resultant macroscopic relationship between pressure and volume cannot be the law enunciated by Boyle. Thus, even if it were the case that nature "preferred" simple laws it would not have the freedom—in a rich universe like ours where there are so many entities and processes, and where there are so many relationships into which physical parameters can enter with one another—to stipulate that all these relationships be simple. Hence the very most a believer in the simplicity of nature could claim is that a small fraction of nature's laws are simple. But then it would seem that the reasonable thing to do, in any individual case, was to expect that the law one is after is, like most laws of nature, a complex one.

The approach favored nowadays, therefore, is to look upon the principle of simplicity as not arising out of any assumptions about the character of nature's laws, but as a principle dictated by some basic methodological propriety. On the surface this may appear a reasonable attitude to take for the objections raised against the ontological justification of the principle of simplicity seem quite convincing as long as one has not too searchingly inquired what exactly is the practical difficulty for the ironing out of which the principle is to be invoked. To get a clear view what kind of basic difficulties one faces when trying to adjudicate among competing hypotheses, we shall consider at some length Jeffreys' views on the subject. His work on simplicity contains invaluable insights.

II

Jeffreys in his *Scientific Inference* asks us to suppose we performed Galileo's inclined plane experiment with the view of obtaining the relationship between the distance s covered by the rolling body and the time t taken to cover that distance. Suppose the following results have been obtained:

t:	0	5	10	15	20	25	30
s:	0	5	20	45	80	125	180

The law which fits these results is given by the equation (a) $5s=t^2$, which is the law enunciated by Galileo. However, the problem arises that there are infinitely many other laws which account equally well for these results.

These laws may be represented by:

(b) $5s = t^2 + t(t-5)(t-10)(t-15)(t-20)(t-25)(t-30)f(t) \ldots$[1]

---

[1] Sometimes it is erroneously said that (b) does not cover all the possible laws which fit the specified data and that there are others which account equally well for them without conforming to schema (b), e.g.,

Where $f(t)$ can take on infinitely many different forms with the only restriction that it should have finite values for $t = 0, 5, 10, 15, 20, 25, 30$.[2] Why then, does the scientist choose equation (a)?

Of the attempts to explain the scientists' approach as an application of a mere methodological rule he considers only the suggestion that the scientist chooses (a) because it is so much more convenient to handle than (b) and Jeffreys rejects this since he takes it for granted that the scientist sincerely believes that (a) is more likely to work than (b), that is, that for hitherto unobserved values of t, by using equation (a) one will obtain the right prediction and not by using any of the equations of form (b).

Jeffreys then points out that even though any one of the infinitely many laws capable of accounting for the observed results may be true and their probabilities add up to one, this does not entail that the *a priori* probability of each one is zero. For, as he explains, it is possible to have a (denumerably) infinite number of finite terms, the sum of which does not exceed one, as for example is the case with the infinite sequence $1/2 + 1/2^2 + 1/2^3 + 1/2^4 + \ldots$. Thus there is nothing wrong with saying that there are (denumerably) infinitely many possible laws which fit the observed results and each one of these has some non-zero prior probability. For this amounts to no more than claiming that the prior probabilities of all the possible laws are represented by the terms of a continually decreasing convergent series whose sum is one. If in addition we also claim that the simpler the law is, the earlier its probability occurs in the series then we have an adequate explanation why the scientist chooses the simplest law.

It is obvious that the most one can claim for Jeffreys' position is that he had succeeded in showing that if we *had* some good reason to claim that simple laws are more probable than complex laws and that their probabilities varied as the terms of an infinite series whose sum converged to unity, then we could arrange all the admissible laws in the order of their decreasing simplicity and be able to assign a specific finite value to the probability of any one of the infinitely many possible laws. I say this is the most he might have achieved because some have seriously questioned whether indeed he has succeeded in devising a satisfactory method of ordering equations uniquely according to their complexity.[3] Be it as it may, Jeffreys has certainly not provided any proof nor did he claim to have provided a proof that there is no other way of justifying the choice of a given hypothesis on the basis of probability considerations and in a way compatible with the common probability calculus but by assigning the terms of a converging infinite series as the values of the prior probability of hypotheses ordered by their decreasing simplicity.

Obvious as this may strike most of us, it does not appear so to everyone. J. C. Harsanyi who discusses Jeffreys' views on probability and simplicity says, for example, the following:

> The rule that simpler hypotheses should be assigned higher a priori probabilities can be justified in various ways. Jeffreys has shown that this rule directly follows from the axioms of the probability calculus if we add the requirement that no possible hypothesis should be allotted zero probability. For the sum of all the probabilities cannot exceed unity; and this is possible only if listing all the infinitely many admissible hypotheses in order of increasing complexity the probabilities assigned to them form a diminishing number series (possibly with a finite number of exceptions).
>
> There is, however, I believe a more fundamental philosophical reason why simpler hypotheses should be assigned higher a priori probabilities. The reason is that they involve *a smaller number of independent assumptions.*[4]

$$\text{(c)}\quad 5s = t^2 + \sin t.\sin(t-5)\ldots.\sin(t-30)\ldots.$$

In fact, however, by judicious choice of $f(t)$, (b) reduces to any proposed form. In particular, by letting

$$f(t) = \prod_{n=0}^{6}\left(\frac{\sin(t-5n)}{t-5n}\right) \text{ for } t \neq 5k \text{ where } k \text{ is an integer and } 0 < k < 6$$

$$\text{and } f(t) = \prod_{n=0}^{k-1}\left(\frac{\sin(t-5n)}{t-5n}\right).\ \prod_{k+1}^{6}\left(\frac{\sin(t-5n)}{t-5n}\right) \text{ for } t = 5k$$

(b) reduces to (c). This point I owe to Drs. A. M. Hasofer and J. C. Burns.

[2] I believe it is fair to say that Jeffreys would agree that $f(t)$ must yet further be restricted so that the resultant variation of s with t conforms to certain commonsense observations known long before Galileo, e.g., that bodies subject to the pull of the earth keep (fairly close) to the same direction throughout their fall (or at any rate throughout the first few hundred yards of their journey).

[3] See R. Harré, "Simplicity as a Criterion of Induction," *Philosophy*, vol. 34 (1959), pp. 229–234; and Robert Ackermann, "A Neglected Proposal Concerning Simplicity," *Philosophy of Science*, vol. 30 (1963), pp. 228–235. The latter shows quite convincingly that Jeffreys has not succeeded to achieve what he set out to do.

[4] J. C. Harsanyi, "Popper's Improbability Criterion for the Choice of Scientific Hypotheses," *Philosophy*, vol. 35 (1960), pp. 309–312.

It is clear, however, that to satisfy Jeffreys' requirements, it is not necessary that we arrange infinitely many admissible hypotheses in order of increasing complexity and assign to each of them a term of a converging series. It is first of all obvious that the same thing can be achieved if we list all the infinitely admissible hypotheses in any other order and assign probabilities to them which correspond to the terms of a convergent number series. In addition the requirement that the sum of all the probabilities should not exceed unity may be reconciled with the fact that no possible hypothesis should be allotted zero probability and furthermore that they are all equiprobable, as long as there are only finitely many admissible hypotheses in any given case. This may be achieved by stipulating that a hypotheses does not qualify as admissible by the mere fact that it satisfies all the experimental results—it also has to meet certain restrictions of form. Then there is yet a third and most important possibility—and this in fact is what I shall attempt to show to be the case—according to which we need not place any implausible restrictions on the number of admissible hypotheses. They may be, as one would naturally assume them to be, non-denumerably infinite. It need not disturb us if the prior probability of any specific hypothesis is zero. What matters is that the "simplicity-conjecture," namely the conjecture that the *kind* of law operating is simple, which initially might have had an extremely low probability, keeps increasing with the increase in the number of experimental data. It also turns out as we shall see these data are compatible but with one of the infinitely many simple laws that were originally admissible, namely with law (a). Consequently the probability that law (a) is true, while prior to experiment it was zero, keeps increasing indefinitely and approaches certainty with the accumulation of experimental results.

It may be of some interest to point out that the claim made in the second paragraph of the quoted passage is not substantiated. There is no reason whatever for saying that to maintain law (b) necessarily requires the making of more assumptions than to maintain law (a). Were we dealing with rival hypotheses which differ from one another in the type of physical mechanism they postulated to be underlying the given phenomenon, Harsanyi's claim might have been relevant. If hypothesis (a) postulated a simple mechanism with a few physical components while hypothesis (b) postulated a complex mechanism in which a large number of factors combined to give rise to the observed phenomenon it would make good sense to say that hypothesis (b) rested on a larger number of independent assumptions than hypothesis (a). But in cases where nothing is said about the mechanism responsible for a given functional relationship and the only question is which mathematical form represents it correctly, it makes no sense to speak about making more independent assumptions by assuming a more complex than a more simple form. In our particular case we either make the single assumption that time and distance vary in accordance with the equation $5s = t^2$, or the alternative single assumption that they vary in a manner represented by the equation $5s = t^2 + t(t-5)$, etc.

III

A solution based on probability considerations can, however, be shown to follow without making any elaborate presuppositions Jeffreys asks us to make. As a first step we must agree that given a finite number of points on a plane, it is not a certainty that a simple curve will pass through them. For my present purposes I regard a curve as simple if and only if the law governing the co-variation of two physical parameters represented by it is regarded as simple by the scientific community. I am not going to give a definition of what is to count a simple law. The point is that no matter what definition we have in mind, or indeed whether we have any definition in mind, we are compelled to agree that there are sets of points, representing parameter couples, through which it is impossible to draw any simple curve. Otherwise we would be forced to say that simple laws and only simple laws operated in any logically possible universe, since given a finite number of observations of the co-variation of any two physical parameters—and we can only be given a finite number of them—a simple law can always account for them.

In reply to this it could perhaps be said that while it would certainly be absurd to assume in all cases of physical parameter-couples that a simple curve may be drawn through a set of points representing their co-variation, we might do so in some selected instances. If so, all we have to ask is whether the co-variation of time and distance of freely falling bodies belongs to this select group? One who answers *Yes*, need go no further; for him the use of the principle to choose the simple law compatible with the data rather than any of its complex rivals is fully justified. But not many shall be happy with such a justification. That is why the problem we set out to solve will seem to most a very real one, namely, how to justify the use of the principle of simplicity in a situation like the one confronting Galileo? Thus we take it that the answer is *No*: We do not assume that a simple curve may be drawn through any number of points representing the co-variation of s and t where the values of one of them has been chosen at random. Having been granted this, we shall also agree then, of course, that if n is the minimum number of such points through which it is no longer a certainty that a simple curve may be drawn, then as we increase the number of those points beyond n, the probability that a simple curve will pass through them keeps decreasing. The only additional assumption we shall use in what follows, and this will be granted at once, is that *some* curve will certainly pass through a finite set of points no matter how large.

Now let C stand for the proposition: A complex curve passes through the points on the paper in front of me.

And let S stand for the proposition: A simple curve passes through the points on the paper in front of me.

And let L stand for the proposition: The law governing the co-variation of time and distance belongs to that *large* group of laws of nature which are represented by a *complex* curve.

And let T stand for the proposition: The law governing the co-variation of time and distance belongs to that particular group of laws of nature which are represented by a *simple* curve.

And let P_n stand for the proposition: There are n points on the paper in front of me and they represent the results of experiment to determine the co-variation of time and distance.

Employing the usual notation where "$p(T/P_nS)$" stands for "the probability of T given that P_n and S" we have by the conjunctive axiom of probability

$$p(T\,.\,S/P_n) = p(T/P_n)\,.\,p(S/T\,.\,P_n) = p(S/P_n)\,.\,p(T/S\,.\,P_n).$$

$$\text{Hence } p(T/S\,.\,P_n) = \frac{p(T/P_n)\,.\,p(S/T\,.\,P_n)}{p(S/P_n)}$$

But since P_n and T together entail S, $p(S/T\,.\,P_n) = 1$. (If it be objected that we have disregarded the possibility of experimental error let S stand for: a simple curve passes through or *sufficiently close to* each one of the points on paper in front of me.)

$$\text{Thus } p(T/S\,.\,P_n) = \frac{p(T/P_n)}{p(S/P_n)}$$

Employing the disjunctive and then the conjunctive axiom of probability we get:

$$\begin{aligned} p(S/P_n) &= p(T\,.\,S/P_n) + p(L\,.\,S/P_n) \\ &= p(T/P_n)\,.\,p(S/T\,.\,P_n) + p(L/P_n)\,.\,p(S/L\,.\,P_n). \end{aligned}$$

Now, since as we have said $p(S/T\,.\,P_n) = 1$, the first term of the R.H.S. reduces to $p(T/P_n)$. The prior probability of L/P_n may be assumed as near unity as we please. The crucial point is that $p(S/L\,.\,P_n)$ keeps decreasing with increasing n as explained. Thus the value of $p(S/P_n)$ keeps approaching $p(T/P_n)$ from which it follows that $p(T/S\,.\,P_n)$ gets closer and closer to unity with the increase of n. We see therefore that while the prior probability that the law we are after belongs to the small set of laws characterized by their simplicity may not be appreciable, it greatly increases after obtaining experimental results and seeing that a simple curve passes through the points representing them. Indeed the conjecture that T, given that S and P_n approaches certainty. But eventually the only law left which is both compatible with T and the experimental results is law (a);[5] hence the probability that law (a) is the true law varies as $p(T/P_n\,.\,S)$.

On the other hand the probability that out of the infinitely many complex equations of form (b) any particular one represents the true law of nature is zero even initially when the probability of L may be taken as near to unity as we please (unless we assumed that the admissible complex equations were merely denumerably infinite and that their probabilities varied as the terms of a converging series). Subsequently, of course, the probability of L itself decreases indefinitely as the probability that the relationship between time and distance is truly represented by $5s = t^2$ steadily increases with our collection of data.

IV

All this may strike us, however, as too good to be true and we may begin to think that we achieved what we did by dividing all curves into two groups just in such manner as to ensure the desired result. Thus the suspicion arises that one could devise an argument which would just as powerfully show that the probability of some particular law represented by a specific equation of form (b) keeps increasing with the accumulation of experimental results on the basis of the division of all curves under a different principle.

---

[5] For a detailed argument supporting this contention see the latter part of this article.

Let us have this objection laid out in detail: Suppose we compile a small list of curves which will include one particular curve of form (b) and a few dozen other curves chosen in any way one may think of—with the only restriction that given a finite number of points on a plane it should not always be a certainty that some curve, the kind of which is on our list, passes through them. Let the curve represented by (a) not be on this small arbitrary list of curves. Now let :

C' stand for the proposition: A curve *not* to be found on our small list of arbitrarily selected curves passes through the points on the paper in front of me.

S' stand for the proposition: A curve *to be* found on our small list of arbitrarily selected curves passes through the points on the paper in front of me.

L' stand for the proposition: The law governing the co-variation of time and distance belongs to that large group of laws which are represented by a curve *not* to be found on our small list of arbitrarily selected curves.

T' stand for the proposition: The law governing the co-variation of time and distance belongs to that particular group of laws which are represented by a curve *to be* found on our small list of arbitrarily selected curves;

then it appears that $p(T'/P_n . S)$ approaches one for exactly the same reason as $p(T/P_n . S)$ approaches one and this completely destroys our previous justification for choosing (a).

The crucial point, however, which has been overlooked in this argument is that even though $p(S'/P_n)$ may be approaching $p(T'/P_n)$, dividing $p(T'/P_n)$ by $p(S'/P_n)$ does not necessarily give us a large number or indeed finite number, if to begin with $p(T'/P_n)$ equals zero! And the scientist will in fact assign zero credibility to T'. After all, there are infinitely many ways in which arbitrary lists may be compiled, and, as one list is as good as the other, it is not unreasonable for the scientist to assign zero probability to the representability of the law we are after, by a curve belonging to a specific arbitrary list. However, the "list" which contains all the simple curves enjoys a slightly privileged status. It is not required that the scientist should believe that all laws are simple, nor even need he be convinced that some laws are simple. It is sufficient that he maintains that there is a finite probability that some of the laws which are likely to come to our attention are simple: in other words the initial probability that T, however small it may be, is not zero. The question what the precise status of this assumption is will be discussed in Section VI.

V

Let me now make explicit the crucial points I hope to have established if the approach outlined in the previous section works.

First, I have been anxious to ensure that it is understood that the strong belief in a typical post-experimental situation that the simple law is immensely more likely to be true than any one of its infinitely many rivals is ultimately based on a far more innocuous assumption than it appeared at first sight—an assumption not involving the belief in the considerable prior likelihood of a simple law operating in a given case. An important aspect of the situation on which I have tried to focus attention is, however, that it is in the nature of things as elementary mathematics show, that our modest initial assumption, of necessity, becomes immensely magnified upon finding that the large number of experimental results obtained, all fit—within a reasonable margin of experimental error—a simple law.

Secondly, it should be clear now why we need not worry about devising a generally applicable criterion restricting the number of admissible laws prior to experiment to ensure the effective functioning of the principle of simplicity, nor, of course, about the defendibility of the criterion chosen. The number of laws which may be regarded as candidates for being operative in any given case need not be restricted prior to experiment. It should not disturb us if their number is an infinite one and non-denumerably so. Our argument is not based on the finite prior probability of *any given law*. It is based on the finite prior probability of T and L—e.g., on the finite probability that the law operative in our case belongs to this or that *kind* of law.

Thirdly, and most importantly, the vexing, if not insoluble problem of how to give a precise definition of what makes a function simpler than another and how to find an adequate method whereby all the different admissible functions may be graded according to their increasing complexity should now be seen as irrelevant to the really significant problem of how the principle of simplicity works. Surprising as it may sound at first, but according to the present explication, the principle achieves its purpose even though we entirely ignore the question of the relative simplicity of functions and make no more than some very general statements as to what makes a function categorically complex. A function may be said to be definitely non-simple if it is too complex to be handled usefully in the context of the conceptual methods available at a given time and only if it has never in the course of the history of science been advanced as representing a true law of nature.

To demonstrate that this is indeed so, let me first answer what on the surface may seem a serious charge against my account, namely, that I have failed to say anything about a situation where a number of simple hypotheses compete against one another. Suppose a variety of hypotheses—each regarded by the scientific community as simple to one degree or another—may account for all the results obtained. We have neither provided any rule as to how to distinguish between their varying degrees of simplicity, nor a basis for justifying our choice of the simplest among these.

What we must realize, however, is that such a situation is, from a philosophical point of view, relatively uninteresting. All we have to do to escape our predicament is to perform some further experiments for such values of the parameters involved, with respect to which the various rival hypotheses yield different predictions. Whatever the results, the wrong hypotheses will then be eliminated. And there must be such values, otherwise the hypotheses would be equivalent and not competing with one another. In the context of the profound and philosophically really important problem to which Jeffreys has drawn our attention, the difficulty is an ineradicable one. Although each additional experiment eliminates an infinite number of hypotheses there are always still infinitely many of them left in the running.

Suppose that in the context of Galileo's experiment we perform additional experiments for times $t_1, t_2, \ldots, t_n$ and we find the results all fitting equation (a). We would then have eliminated all those (b) type equations in which $f(t)$ does not contain the term $(t-t_1)\ (t-t_2) \ldots (t-t_n)$, but all the other equations in which $f(t)$ does contain this term, and the number of such (b) type equations is infinite, remain to compete against (a). There is no escape from this predicament but by the use of the argument that (a), even though it is only one of infinitely many equations which are equally well satisfied by the experimental data, is immensely more likely to represent the true law operating than any of its rivals for the reasons elaborated here.

Now it should be possible to see clearly why we need nothing more precise than the kind of statement I made in attempting to describe what constitutes a categorically complex function. That statement, to be sure, does not provide a secure enough basis upon which to determine with certainty, whether in the situation described by Jeffreys, we are to regard equation (a) as the sole simple function compatible with the experimental data or perhaps some functions of type (b) may too be so regarded. This, however, need not worry us very much. Some may feel that (a) is the only function which is not categorically complex. According to them, good reason has been provided for preferring (a) to any of the other infinitely many functions.

Suppose, however, that there are others who think that some of the equations of form (b) with a simple kind of $f(t)$ are also not to be regarded as categorically complex. They indeed would not have any good reason to adopt a particular function as the one representing the law of terrestial free fall *at this stage of the experiment.* What they would have to do, therefore, is to obtain some more experimental data to eliminate more hypotheses until a stage is reached where only one of the surviving functions is simple and the remainder categorically complex.

It is of course not a certainty that such stage will ever be reached, or if reached, it will be permanently maintained. It may well happen that all simple functions become casualties of further experimentation. In that case, this particular line of investigation will have to be abandoned. Nothing may be hoped from continuing to observe the covariation of time and distance since there is no longer any basis upon which to adjudicate among the competing hypotheses. The law of the covariation of time and distance may then be hoped to become disclosed only as a consequence of other laws simple enough to be amenable to direct investigation and discovery.

VI

According to the present suggestion, then, the ontological justification of the principle of simplicity need not be based on anything more than the claim that the initial probability in any given case that the law we are after is not zero. It is still a pertinent question which one is entitled to ask: What is the justification of this claim?

First, we should take notice of the fact that the positive arguments mentioned in Section I do not apply against the present justification. That there is no empirical evidence and could be no empirical evidence against the claim on which we based our justification is only too evident, since all we insist upon now before we apply the term "simple" to a function, that its kind has (because of its forbidding complexity) never throughout the history of science been advanced as representing a true law of nature. But it must be obvious also that the second, logical argument does not apply either. We may readily admit that prior to experiment the probability that any given functional relationship is a very complex one is vastly greater than that it is manageably simple. This, however, does not interfere with the proper functioning of the principle which requires no more than that T should not equal zero initially for then it keeps increasing indefinitely with the number of experiments it successfully survives.

Some may wish to go further and maintain that not only are there no positive arguments against assigning non-zero probability to T, but that there are arguments why we should definitely not assign zero probability to it. A believer in the intelligibility of nature, according to whom at least some, if not all, the laws of nature are discoverable by humans, must postulate that finite prior probability attaches to the conjecture that a given law belongs to *some* specific kind, otherwise, as we have seen, no amount of evidence can provide the slightest support to any functional law. The only question then is which particular kind of function be accorded preferential status. Believing in the intelligibility of nature entails the belief that among the laws held at the moment to be true, there is a considerable proportion which is actually true. If so, it seems reasonable to assign finite probability to the event that the law presently under investigation is of the same kind as those already known to us. But the one thing one could safely say about all the laws already known to us to have in common is that they are not categorically complex in the sense earlier defined. As long as no one suggests some other feature that is common to all the laws hitherto entertained, finite probability attached to the conjecture that the desired function is a member of the class of functions characterized by their simplicity, but not to the conjecture that it is a member of some other class.

Others who subscribe to the principle of the intelligibility of nature may wish to go further and argue that a finite proportion of all the laws operating in the universe must be simple for if this were not so, the resulting phenomena would be of such frightful complexity that no simple regularities would be anywhere apparent to provide any foothold upon which to begin a systematic and rational investigation of nature. Yet others would be prepared to concede that the group of laws which are simple may very likely amount to a vanishingly minute fraction of all the laws which govern the universe. The reason why we nevertheless do not assign vanishingly small initial probability to the assumption that a given law is simple is that the very simplicity of a law lends it a unique advantage which raises the likelihood of its discovery to a level which is out of proportion with the distribution of simple laws in nature. A simple law attracts more readily the attention and secures more easily the comprehension of the scientist.

It is conceivable, however, that valid objections can be raised against all these arguments. Some might be of the opinion that in general any attempt to justify an assumption of such fundamental nature as the one that the initial probability of T is not zero must of necessity fail. It must be recognized that ultimately science rests on a number of unprovable assumptions. What one may demand with respect to such assumptions is that it be shown that without them rational inquiry is impossible; that they be kept at a minimum and yet be able to do the job expected from them; that they should not be counter-intuitive and that there be no other positive arguments against their credibility. The assumption that nature is simple in the sense here explicated seems to satisfy these requirements.

Received April 14, 1966

The Australian National University
and
The Minnesota Center for the Philosophy of Science

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 2, April 1967

VIII. HUME'S BUNDLE THEORY OF THE SELF: A LIMITED DEFENSE

NELSON PIKE

THE secondary literature on Hume's *Treatise of Human Nature* contains two major criticisms of the bundle theory of the self. The first centers on the problem of specifying a criterion by which perceptions can be grouped together into individual bundles. Given that the mind is simply a collection of perceptions (as Hume said), by what principle shall we distinguish between the perceptions making up your mind and the perceptions making up mine? Providing an answer to this question is often taken to be the crucial problem confronting any advocate of the bundle theory of the self. It is sometimes thought to be the problem Hume tried (unsuccessfully) to solve in the section of the *Treatise* entitled "Of Personal Identity."[1] The second major criticism of Hume's bundle theory of the self involves the claim that the theory is inconsistent with a certain line of thinking evident in almost every section of the *Treatise*. In "Of Personal Identity," for example, Hume claims to "look within" and discover only a series of perceptions. What is it that looks within and makes this discovery? Surely a collection of perceptions could not perform this task. In his discussion of causation, to take another case, Hume says that the mind remembers that perceptions of a certain sort have been followed in the past by perceptions of another sort. What is this *mind* that remembers perceptions? Surely it could not be the collection of perceptions itself. This same sort of question could be raised with respect to a range of other comments in the *Treatise*. Hume constantly speaks of a mind that not only observes and remembers perceptions, but which unites, confuses, feigns, and believes things about perceptions. Though Hume officially denied the existence of a Transcendent Ego, i.e., a self other than a mere collection of perceptions, the notion of just such a self seems to work importantly in his thinking. In addition to the bundle of perceptions that is supposed to constitute the mind, Hume seems to have presupposed the existence of a self that *deals* with perceptions in a wide variety of ways.

It is clear, I think, that the criticisms just outlined are distinct. Suppose Hume had provided an adequate criterion by which to divide perceptions into individual bundles. We could still ask whether he presupposed the existence of a Transcendent Ego when he claimed to look within and discover only a series of perceptions. Conversely, suppose we agreed that looking within does not require a self other than a series of perceptions. We should still want to inquire about the principle by which perceptions can be divided into individual bundles.

I shall here attempt to defend Hume against the second criticism outlined above. I shall not deal with the problem of providing an adequate criterion by which to divide perceptions into individual bundles. For purposes of discussion, I shall assume that this latter is not a problem—or, better, I shall assume that this problem is not insuperable: it is simply awaiting solution.[2] My remarks will be divided into three sections. In the first, I shall make some preliminary observations about the purpose of Hume's bundle theory of the self. In the second I shall address myself to the specific version of the second criticism to be found in D. G. C. Macnabb's commentary on the *Treatise*. And in the third I shall attempt to show how the defense given in the second section of the paper can be applied to versions of the same sort of attack other than Macnabb's.

---

[1] Cf. for example, A. H. Basson, *David Hume* (Baltimore, Penguin Books, 1958), p. 128.

[2] In the *Philosophy of Logical Atomism*, Russell takes much the same attitude toward this problem. Russell writes: "You can collect a whole set of experiences into one string as all belonging to you, and similarly other people's experiences can be collected together as all belonging to them by relations that actually are observable and without assuming the existence of a persistent ego. It does not matter in the least to what we are concerned with, what exactly is the given empirical relation between two experiences that makes us say, 'these are two experiences of the same person.' It does not matter precisely what this relation is, because the logical formula for the construction of the person is the same whatever the relation may be, and because the mere fact that you can know that two experiences belong to the same person proves that there is such an empirical relation to be ascertained by analysis." (*Logic and Knowledge*, ed. by R. C. Marsh [London, Allen and Unwin, 1956], pp. 276–277.)

I

According to John Locke, words are the ". . . sensible marks of ideas, and the ideas they stand for are their proper and immediate signification."[3] This doctrine ties in with Locke's view that language is used primarily to communicate ideas. Thus a given word (or sound) that does not stand for some specific idea can not be assigned a proper signification or use. And, it would also seem to follow, the analysis of any given word will consist of an examination of the idea for which that word stands.[4] Bishop Berkeley disagreed with this view. Berkeley held that a word may have meaning even if it does not stand for an idea. He offered the terms "spirit" and "will" as illustrations of this claim. Berkeley also disagreed with Locke's view that the chief use of language is to communicate ideas. He said that words are often used to arouse emotions and that this need not involve the communication of ideas.[5] In Book I of the *Treatise,* Hume makes use of a theory of meaning very much like Locke's. According to Hume, ideas are "annexed" to words. If one is to discover whether a given term has meaning and, if so, what meaning it has, one must discover whether there is an idea "annexed" to the word and, if so, what idea it is. To analyze the meaning of a given word is to analyze the idea "annexed" to it. Further, for Hume, to analyze a given idea is to identify the impression(s) (i.e., feelings or sensations) from which that idea is derived. Thus, for Hume, to analyze the meaning of a given word is to identify that impression (or those impressions) from which the idea "annexed" to the word is derived. If there are no impressions, there is no idea and, accordingly, no meaning. If there is meaning, we find it by uncovering the impressions that generated the "annexed" idea. These methodological doctrines are set out in any number of passages in the *Treatise,* but the following passage from the *Enquiry* brings most of them together into compact form:

> All ideas, especially abstract ones, are naturally faint and obscure: the mind has but a slender hold of them: they are apt to be confounded with other resembling ideas; and when we have often employed a term, though without a distinct meaning, we are apt to imagine it has a determinate idea annexed to it. On the contrary, all impressions, that is all sensations, either outward or inward, are strong and vivid: the limits between them are more exactly determined: nor is it easy to fall into any error or mistake with regard to them. When we entertain therefore, any suspicion that a philosophical term is employed without meaning or idea (as is frequent) we need but inquire, *from what impression is that supposed idea derived?* And if it be impossible to assign any, this will serve to confirm our suspicion.[6]

Turning now to the topic of the self, Hume assumes that our idea of *mind* is derived from introspective observation, i.e., from "inward" sensations (to use the phrase introduced in the passage just quoted from the *Enquiry*). On Hume's account, what we perceive *via* "inward" perceptions is a series of numerically distinct perceptions. Hume concludes:

> The mind is a kind of theatre where several percep, tions successively make their appearance; pass, re-pass, glide away, and mingle in an infinite variety of postures and situations. . . . The comparison of the theatre must not mislead us. They are the successive perceptions only that constitute the mind.[7]

Hume's conclusion is here stated in the material mode of speech. It is presented as a theory about the make-up of the mind. But the same conclusion is more enlighteningly stated in the *Abstract to the Treatise*. Here Hume writes:

> As our idea of any body, a peach, for instance, is only that of a particular taste, color, figure, size, consistence, etc. So our *idea* of any mind is only that of particular perceptions, without the notion of anything we call substance, either simple or compound.[8]

The point is that the bundle theory of the self is offered by Hume as an analysis of the *idea* of mind. It is offered as a meaning-analysis of the word

---

[3] *Essay,* Bk. III., ch. II (opening sentence).

[4] *Essay,* Bk. III., chs. I–III.

[5] Cf. paras. 18–20 of the Introduction to the *Principles,* para. 27 of the *Principles,* and para. 14 of the seventh dialogue of the *Alcephron.*

[6] *An Enquiry Concerning Human Understanding* (La Salle, Open Court, 1952), p. 20. Cf. also *A Treatise of Human Nature,* ed. L. A. Selby-Bigge (Oxford, The Clarendon Press, 1958), pp. 33, 48–49 and 74–75; the *Enquiry,* pp. 66–67; and the *Abstract to A Treatise of Human Nature* (Cambridge, University Press, 1938), pp. 10–11.

[7] *Treatise,* p. 253.

[8] *Abstract,* p. 25. Italics are mine.

"mind."[9] At this point I must diverge briefly to consider an objection.

Here and there in the secondary literature on the *Treatise* it is suggested that Hume *denied* that we have an idea of mind. Of course, if this were a correct interpretation of the text, what I have just said concerning the point of Hume's bundle theory of the self would have to be withdrawn. It would be absurd to suppose that Hume offered an analysis of an idea he thought not to exist.

I have two reasons for thinking that this interpretation of Hume is mistaken. First, Hume often makes mention of the idea of mind. The passage just quoted from the *Abstract* is a case in point. And if we turn to Book II of the *Treatise* we can find passages in which Hume is quite clear on the topic.

> 'Tis evident, that the idea, or rather the impression of ourselves is always intimately present with us, and that our consciousness gives us so lively a conception of our own person, that 'tis not possible to imagine that anything can in this particular go beyond it.[10]

And again:

> The idea of ourselves is always intimately present to us and conveys a sensible degree of vivacity to the idea of any other object, to which we are related. This lively idea changes by degrees into a real impression; these two kinds of perceptions being in a great measure the same, and differing only in their degree of force and vivacity.[11]

In these passages, to be sure, Hume does not explicitly mention the idea of *mind.* He speaks only of the idea of *ourselves* or our own *person.* But, as is agreed by most scholars of Hume's writings, when Hume uses the term "self" or "person," he generally means to be referring only to the mind, not (as we might say) to the *whole* person.[12]

Of course, one might argue at this point that while Hume affirms the existence of an idea of mind in Book II of the *Treatise,* he denies its existence in Book I. On the opening page of his commentary on Hume's philosophy, Norman Kemp Smith asks the following question:

> Why is it that in Book I of the *Treatise,* the existence of an impression of the self is explicitly denied, while his theory of the "indirect" passions propounded at length in Book II is made to rest on the assumption that we do in fact experience an impression of the self, and that this impression is ever-present to us?[13]

This brings me to my second reason for rejecting the above interpretation of Hume concerning the existence of the idea of mind.

In the first two paragraphs of "Personal Identity," Hume considers a view that he ascribes to "some philosophers." According to some philosophers, Hume says, the self is a simple substance that has perceptions. It is not simply the collection of perceptions itself. Unfortunately for these philosophers, Hume continues, we have no idea of the self "*after the manner it is here explained*" (italics are mine). We have no impression of a self-substance that has perceptions, and, accordingly, "we have no *such* idea" (italics are mine). I think it has been these remarks which have led Smith (and others) to suppose that Hume explicitly denied the existence of an idea of self in Book I. But if one looks closely at these paragraphs, I think it is plain that they contain no such denial. Hume says that we have no idea of a *certain kind* of mind, viz., a mind-substance which has perceptions. He does not say that we have *no* idea of mind. We do have an idea of mind—as the passages just quoted from Book II make clear. Under analysis we discover that this idea is not the idea of a mental-substance; it is, rather, the idea of a bundle of perceptions. There is no inconsistency of the sort alleged by Smith. Hume rejects a certain philosophical *analysis* of our idea of mind, but he does not deny that there is an idea of mind to be analyzed.

---

[9] Among contemporary empiricists, the bundle theory of "mind" is usually formulated explicitly as a theory about the meaning of the single term "mind" (cf., e.g., C. I. Lewis, *Mind and the World Order* [New York, Scribner, 1929], p. 421) or, in more sophisticated fashion, as a theory about the empirical meaning of whole sentences about the mind (cf., e.g., A. J. Ayer, *Language Truth and Logic* [New York, Dover Press, 1946], pp. 123–124). In this connection, it is interesting to note that while Bishop Berkeley spoke of the self in his published works as a substance which "has" perceptions (e.g., *Works,* vol. II, ed. A. A. Luce and T. E. Jessop [London, Nelson and Sons, Ltd., 1948], pp. 232–234) in his *Notebook A* he entertained the bundle theory of the self. Further, Berkeley's formulation of this theory (more obviously than Hume's) casts it as a theory about the meaning of the word "mind." Berkeley writes: "The very existence of ideas constitutes the soul. Consciousness, perceptions, existence of ideas seem to be all one. Consult, ransack your understanding. What find you there besides several perceptions or thoughts. *What mean you by the word mind?* . . . To mean a thing you do not perceive is a contradiction. We are in this matter strangely abused by the words. Mind is a congeries of perceptions. Take away the perceptions and you take away the mind. Put the perceptions and you put the mind. Say you the mind is not the perceptions, but the thing which perceives. I answer you are abused by the words . . . These are vague, empty words without meaning." (*Works,* vol. 1, p. 72. Italics are mine. I have taken the liberty of re-punctuating these passages. Berkeley's punctuation makes them hard to read.)

[10] *Treatise,* p. 317. The sixth paragraph of the section of the *Treatise* (Bk. I) entitled "Scepticism with Regard to the Senses," also contains mention of our "ideas of self and person."

[11] *Treatise,* p. 354.

[12] This point is made nicely by Terence Penelhum, "Hume on Personal Identity," *The Philosophical Review,* vol. 64 (1955), pp. 575–586.

[13] *The Philosophy of David Hume* (London, Macmillan, 1960), p.v.

II

Consider the statement: "I see a chair." I shall call this statement S. It is sometimes supposed that the whole point of a theory of the self (such as Huem's) is to identify the object referred to by the pronoun 'I' in sentences such as S. But whatever may be true of some theories of the self, this is not an adequate description of the theory put forward by Hume. Hume advanced his theory as an analysis of "*mind*." He had nothing directly to say about the meaning of the pronoun 'I'. Yet it is clear, I think, that sentences such as S say something about my mind. They tell us something about what is going on *in* my mind. S reports that the visual perception of a chair is presently occurring in my mind. This is not to say that *my mind* is seeing a chair. *I* am seeing the chair. But if I am seeing a chair, the visual perception of a chair is presently occurring in my mind. One of the messages of "Scepticism with Regard to the Senses" is that "I see a chair" is reducible to the statement "The visual perception of a chair is presently occurring in my mind."

Let us then assume (in accordance with Hume's intentions) that S reports the presence of a visual chair-perception in my mind. Given the bundle-analysis of "mind," the phrase "my mind" must be understood as designating that special collection of perceptions which is *mine*. And which collection is that? The collection of perceptions belonging to some pure ego—me? Surely not. "My mind," considered as a meaningful sound produced by this body (the body sitting in a certain chair at a certain time, etc.) designates a certain collection of perceptions ϕ. It is the same collection designated by the phrase "Pike's mind" when others are speaking. Had we an adequate criterion by which to divide perceptions into individual bundles, we could now go on to provide a unique description of the collection ϕ—a description that would make no mention of a pure-ego that "has" the perceptions constituting ϕ. The collection ϕ is the collection of perceptions whose members are related to one another by the special relation R; or, the collection ϕ is the collection of perceptions whose members share a common introspectable feature F; or some such. Of course, we do not actually have such a criterion. But, as pointed out above, I am putting this problem aside for the present. The point is that for Hume, "my mind" must designate a certain collection of perceptions that is specifiable without mention of anything other than perceptions, their qualities or relations. Thus for Hume, "I see a chair" reduced to "The visual perception of a chair is presently occurring in my mind" and this, in turn, reduces to "The visual perception of a chair is presently in the collection of perceptions ϕ" (where the criterion for identifying ϕ requires mention of nothing other than perceptions, their qualities and relations).

We can see, I think, that one result of Hume's analysis of "mind" is that it provides a specific analysis of the "in" in "in my mind." It might be thought that "in" is properly a special term occurring in the phrase "in my mind" in a metaphorical sense. I can find no reason to think this is true. "In" is not always used to signify spatial place. It has plenty of non-metaphorical uses that have nothing to do with space (e.g., "*in* a hurry," "*in* his name," "*in* the mood"). I think a more enlightening way to look at the matter is to think of "in my mind" as a pre-analytical phrase requiring philosophical attention. With this understanding, what Hume is suggesting is that the "in" in question is the "in" of membership (e.g., "I am *in* that club"). The visual perception of a chair is a present *member of* the collection of perceptions ϕ.

Let us now consider the following passage from the closing page of D. G. C. Macnabb's discussion of Hume's theory of knowledge. Macnabb writes as follows:

> A difficulty has been formulated by someone in a hexametrical criticism of Hume. "How can a series of conscious states be aware of itself as a series?" But when I consider this hexameter, I see that the difficulty has been wrongly put. It is not the series as a whole which is required to be aware of itself. We are not self-conscious all the time. We are self-conscious at certain times. It is some of the members of the series that must be aware of themselves as members of the series. Just as when a nation is said to be conscious of itself as a nation, it is really the members of the

> nation, or some of them, who are aware of themselves as members of the nation. But still, what is it for a perception to be aware of itself as a member of that relational unity of perceptions we call mind? I do not know the answer to this question.[14]

"How can a series of conscious states be aware of itself as a series?" The implication is that if anything is aware of a series of conscious states, it must be something other than the series itself. A series of conscious states cannot be aware of itself as a series. Surely, this last observation is correct. A series of conscious states can not be *aware of* anything at all. *I* can be aware of things. But a series of conscious states is not the *kind of thing* that can be aware of something—itself or anything else.

But I can find no special problem here for Hume. Had Hume offered his bundle-analysis as an account of the meaning of 'I' (in such statements as "I am aware of myself as a series of perceptions") he would now be forced to adopt the erroneous view that a bundle of perceptions can be aware of itself as a bundle. But Hume analyzed "mind," not 'I'. Thus if we are to discover what Hume would say about the meaning of the statement: "I am aware of myself as a series of perceptions," we must first "translate" it (as it were) into a statement making explicit mention of my mind. Following the pattern suggested above, I think Hume would translate this statement as follows: "An awareness of myself as a series of perceptions is presently occurring in my mind." This last statement, given Hume's bundle-analysis of "mind," would mean: "An awareness of myself as a series of perceptions ϕ is presently occurring within the series of perceptions ϕ." Of course this last still contains the term "myself" (i.e., my "mind") so the bundle-analysis must be applied again. The final analysis is then: "An awareness of the series of perceptions ϕ (in which ϕ is presented as a series of perceptions) is presently occurring within the series of perceptions ϕ." This last statement is awkward, but it is intelligible. A series of conscious states can not be *aware of* itself as a series. But a series of conscious states might *contain* an awareness which is *of* itself presented as a series. Analogy: A collection of pictures cannot *paint* a picture which is of itself pictured as a collection. But a collection of pictures might *contain* a picture which is of itself pictured as a collection.

Macnabb says that the real problem is yet to come. Hume need not say that the whole series of perceptions is aware of itself as a series. He need only say that some members of the series are aware of themselves as members of the series. The real problem, says Macnabb, is how a given perception can be aware of itself as a member of a certain series of perceptions. Again, the implication is that if anything is aware of a perception as a member of a series of perceptions, it must be something other than a perception. As above, I can only agree with this observation. A perception can not be aware of itself as a member of a series of perceptions any more than a series of perceptions can be aware of itself as a series. *I* can be aware of things. But a perception (like a series of perceptions) is not the *kind* of thing that can be aware of anything at all.

Again, I can find no problem here for Hume. *I* am aware of a certain perception as being a member of a certain series of perceptions. Translating this statement into one making an explicit mention of my mind, we get: "An awareness of a certain perception as being a member of a certain series of perceptions, is occurring in my mind." And, given Hume's bundle-analysis of "mind," this last means: "An awareness of a certain perception as being a member of a certain series of perceptions ϕ is occurring within that series of perceptions (i.e., within ϕ)." I can see no difficulty here. This last statement might mean: "A certain series of perceptions ϕ includes a member *A* which is of another member *B* standing in relation to some other members *C, D, E,* etc." Or, alternatively, it might mean: "A certain series of perceptions ϕ includes a member *A* which is of itself (i.e., of *A*) standing in relation to some other members *B, C, D, E,* etc." On neither interpretation does it affirm that a perception is *aware of* a perception (itself or another perception) or that it is *aware of* the series of which it is a member. On both readings, it says only that the collection of perceptions ϕ *includes* a member which is *of* a perception (itself or otherwise) standing in relation to some other perceptions. This seems intelligible. My picture-collection includes a picture which is of a picture (itself or another picture) hanging next to some other pictures in the collection.

III

I want now to generalize this defense of Hume. But before I do, I shall consider one more specific version of the kind of charge with which I have been dealing.

---

[14] *David Hume: His Theory of Knowledge and Morality* (London, Hutchinson, 1951), pp. 251–252.

In *Hume's Intentions*, J. A. Passmore writes as follows:

> The fundamental problem (in Hume's discussion of personal identity) is what it is which confuses succession with identity; and equally, what it is which comes to recognize that succession *has been* confused with identity? . . . If all that happened is that a series of very similar (or causally linked) perceptions succeed one another, there is no possible way in which this series of itself could generate the fiction of personal identity. Nor the fiction once generated, could this series ever reveal its fictional character. Both the original fiction and the discovery that it is a fiction are possible only if there is something which is at first misled by, and then, after consideration, can discover that it was misled by, a series of similar perceptions. . . . (It is) at this point . . . that Hume's psychological positivism breaks down.[15]

Hume held (in "Of Personal Identity") that the perceptions making up the mind are separate and distinct entities—he called them "substances."[16] But, according to Hume, the common man believes, or has a propensity to believe, that these separate and distinct entities are "moments in" or aspects of a single, continuous, uninterrupted perception (i.e., a single, continuous, uninterrupted experience). The common man has a propensity to ". . . ascribe an identity to these successive perceptions,"[17] i.e., to ". . . suppose the whole train of perceptions to be united by identity,"[18] i.e., to ". . . run the several different perceptions into one, and make them lose their characters of distinction and difference which are essential to them."[19] Of course, Hume thought this belief false. Though it is to be distinguished from the belief held by "some philosophers" concerning the existence of a self-substance (Hume does not ascribe the belief in self-substance to the common man), it, like the belief in self-substance, is mistaken. Further, Hume clearly supposed that careful inspection of the contents of mind will establish this fact. Thus, according to Hume, *I* believe (or have a propensity to believe) that successive perceptions are "moments in" a single, continuous, uninterrupted experience. And *I* discover that this belief is in error.

Let's look at these last two statements. Hume would say that the first affirms that a certain belief-perception (i.e., a certain idea which is especially "strong" and "lively") is occurring *in my mind*. The second (I think Hume would say) affirms that a second perception (call it a "realization-of-error-perception") also occurs *in my mind*. Given Hume's bundle-analysis of "mind," these last mean only that certain kinds of perceptions (a belief-perception in the first case, and a "realization-of-error-perception" in the second) are included in a certain series of perceptions ϕ. There may be difficulties with this view. For example, there may be difficulties with Hume's claim that a belief is simply an idea which is especially "strong" and "lively." But so far as I can see, there is no difficulty of the sort urged by Passmore. Neither of these claims presuppose a self other than a series of perceptions—one which confuses succession with identity and later discovers that it has made a mistake. These claims entail only that certain kinds of perceptions are included in a certain collection of perceptions.

I think our findings can now be generalized. According to Hume, a statement of the form "I see a chair" (where the object perceived is a physical object) is reducible to a statement of the form "A visual chair-perception is occurring with a certain bundle of perceptions ϕ." When the object perceived is a physical object, the perception verb (in this case "see") is ultimately excludable. Statements containing such verbs are translatable into statements mentioning only the presence of a certain perception (i.e., a certain "substance") within a certain collection of perceptions. So much, I think, is clear in Hume's texts and has been generally recognized by most commentators on Hume's philosophy. I think Hume intended the same to be true with regard to such statements as "I observe a chair-perception"—where the object perceived is not a physical object but is, instead, an introspectable mental state. "I observe a chair-perception" means "An ("inward") perception of a chair-perception is occurring in a certain bundle of perceptions ϕ." As a general remark, Hume seems to have thought that statements about the *activities* of mind (such as observing, remembering, believing, etc.), whether these activities are activities with respect to physical objects or mental states (i.e., whether it is physical objects or mental states which are being observed, remembered, etc.) are ultimately reducible to statements affirming the presence of certain specialized perceptions in the collection of perceptions making up the mind. Mental verbs, in other words, are not to be included in the finished mental vocabulary.

---

[15] Cambridge (1952), pp. 82–83.
[16] *Treatise*, p. 233.
[17] *Treatise*, p. 253.
[18] *Treatise*, p. 269.
[19] *Treatise*, p. 259.

When we turn to Hume texts, we often find him saying that *the mind* perceives, believes, remembers, etc., things. Such statements suggest that there is a mind which *does* things. When the objects with respect to which the mind (allegedly) does something are perceptions (i.e., mental states) critics have charged that Hume has supposed the existence of a mind other than a mere collection of perceptions. It is hard to see how a mere collection of perceptions could be that which perceives, believes things about, remembers, etc., the very perceptions of which it is composed. Of course, it is equally hard to see how a mere collection of perceptions could be that which perceives, believes things about, remembers, etc., physical objects too—though Hume's critics (so far as I know) have not sensed the need for a Transcendent Ego in these cases. The key to the problem, I think, is that for Hume, the mind does not *do* anything—it *includes* things. Statements containing mental verbs (whether the objects receiving the action of such verbs be physical objects or mental states) are used by Hume as conveniences—manners of speech. We get to the real meaning of such statements *via* a Humean analysis. In analysis, statements mentioning the *activities* of mind are replaced by statements affirming only the presence of certain mental *substances* (perceptions) within a certain collection. Hume presupposes no Transcendent Ego. Given his analysis of "mind" there is nothing for such an Ego to *do*.

Cornell University

Received February 7, 1966

American Philosophical Quarterly
Volume 4, Number 2, April 1967

IX. THE INNER MAN

GARETH B. MATTHEWS[*]

> . . . there is someone within who can see through the eyes. So when sometimes you think within you—the inhabitant having turned inwardly aside—you do not see what is before your eyes. For the windows are open in vain when he who gives heed through them is absent. It is therefore not the eyes which see, but someone who sees through the eyes. Awake him, arouse him![1]

Of all philosophers St. Augustine is perhaps the most enamored of talking about the mind as a man inside. Perhaps one reason for this is that the expression "*homo interior*" has for him a certain theological sanction; St. Paul uses its Greek equivalent (*ὁ ἔσω ἄνθρωπος*) in three of his epistles.[2] But perhaps just as influential on Augustine's thinking, and philosophically rather more interesting, are two inner-man ways of speaking characteristic of Biblical writers generally, viz., (1) the personification of parts of the body ("The eye of the adulterer also waits for the twilight, saying, 'No eye will see me' "—*Job* 24:15; "The ear of the wise seeks knowledge"—*Proverbs* 18:15) and (2) the report of what a man thinks, hopes, fears, and desires as what he says or does in his heart ("Then I said in my heart, 'What befalls the fool will befall me also . . .' "—*Ecclesiastes* 2:15; "You have heard that it was said, 'You shall not commit adultery.' But I say to you that everyone who looks at a woman lustfully has already committed adultery with her in his heart."—*Matthew* 5:27–8).

By supposing that the inner man may be said to have eyes, ears, a mouth, etc.,[3] and by supposing that whatever a man says or does in his heart is done "inside," by the inner man,[4] Augustine links these two Biblical ways of speaking to the general notion that the mind is a little man inside.

Typically Augustine speaks of remembering, imagining, visualizing, conceiving, and understanding as cases of seeing something with the inner eye.[5] He describes what a man is thinking on a particular occasion, as well as the intention he forms to do something, as the speech of an inner voice, or, equivalently, as what a man says in his heart.[6] He conceives the desire to do something as doing it already in one's heart.[7] And he distinguishes the man who acts with a pure motive from his Pharasaical neighbor by saying that, though both, for example, give alms outwardly, only the man of pure motive also gives alms inwardly.[8] In short, the family of inner-man locutions provides for Augustine a connected way of conceiving mental functions and narrating psychological episodes.

I

Augustine's conception of the function of words like "inner," "inwardly," and "in the heart" is simple and direct. He expresses it here:

---

[*] I wish to thank Professor Peter Geach for reading through this paper carefully and criticizing it. His comments have made me wiser even where I have not been able to make the paper any better.

[1] Augustine, *Sermo* 126.2.3. (Translations from Augustine in this paper are my own, though I have freely consulted existing translations.)

[2] "For I delight in the law of God in my *inner man*, but I see in my members another law at war with the law of my *mind* and making me captive to the law of sin which dwells in my members." (*Romans* 7:22–3; emphasis mine.) Other occurrences of the expression are to be found at 2 *Corinthians* 4:16 and *Ephesians* 3:16. For an example of Augustine's concern with what "inner man" means in the Pauline epistles see Augustine's *Contra faustum manichaeum* 24:2. (Quotations from the Bible in this paper follow the Revised Standard Version, with occasional alterations.)

[3] ". . . our very own mind, that is, the inner man, to whose uniform knowing diverse things are announced through the five, as it were, messengers of the body . . . sees the light . . . hears the word . . . catches the scent . . . drinks of the fountain. . . . and enjoys the feel . . . Nor are these different things, but one understanding that is expressed by the names of so many senses" (*In joannis evangelium* 99.4).

[4] "The inner man has an inner mouth . . . [and] the things that proceed from this mouth come from the heart" (*De continentia* 2.4).

[5] See especially *De genesi ad litteram*, Bk. XII, on the "three kinds of vision."

[6] See, e.g., *Sermos* 187, 225 and 288 and *De trinitate* 7.7.12.

[7] See, e.g., *Sermo* 179.7.7.

[8] See, e.g., *De sermone domini in monte* 2.2.9.

> . . . a man is not just a body, or just a soul, but a being made up of both body and soul. . . . The soul is not the whole man, but the better part of a man; the body is not the whole, but the inferior part of a man. *When both are joined together they have the name "man," which, however, they do not either one lose when we speak of them singly.* For who is prohibited from saying, in ordinary language, "That man is dead and is now in peace or in torment," though this can be said only of the soul; or "That man is buried in that place or in that," though this cannot be understood except as referring to the body alone? Will they say that Holy Scripture follows no such usage? On the contrary it so thoroughly adopts it, that even when a man is alive, and his body and soul are joined together, it calls each of them singly by the name "man," speaking of the soul as the "inner man" and the body as the "outer man," as if there were two men, although both together are one man. (*De civitate dei* 13.24.2; my italics.)

Augustine's idea is that the function of "inner" in the expression "inner man" is to make specific what otherwise would have to be surmised from the context, viz., that on this occasion one is using "man" to refer to only that part of a man which is separable from the body at death, viz., his mind or soul.[9] And the function of "outer" in the expression "outer man" is correspondingly to make explicit a reference to the body alone. In view of the fact that Augustine treats inner-man locutions in a connected way this means, first of all, that on his view inner seeing, inner speaking, inner smiling, inner giving, etc.—since they are all activities of the inner man—should be considered activities of an entity distinct and separable from the body. It means, secondly, that whenever a man is said to do or say something inwardly (or in his heart), this, too, should be thought of as the action of an entity distinct and separable from the body.

If Augustine were right about all this, then only a metaphysical dualist could allow himself the use of inner-man locutions without being inconsistent. For to say of someone that he wept inwardly, or that he could see his old sweetheart in his mind's eye, would be to say something that entails metaphysical dualism. What reason is there for thinking that this is so?

II

When the outer man has been deserted by its inhabitant, Augustine tells us, "there is no one to move the steps to walk, the hands to work, the senses to perceive" (*Sermo* 173.3.3). What has left the body is something that, when present, "rules the body, moves the limbs, guides the senses, prepares thoughts, puts forth actions and obtains images of innumerable things" (*Enarrationes in psalmos* 145.4).

With this causal picture of the relation between mind and body in view, one might expect that the locutions of inner seeing, inner speaking, inner smiling (etc.) would offer a basis for causal explanation of certain abilities that human beings have. Indeed Augustine does seem to suppose that the explanation of our ability to visualize the face of an absent friend rests on the fact that we have a mental picture on file that we can take a look at from time to time.[10] And one might say that, the reason I can, in general, be expected to know what I am thinking, is that my inner ear is in very close proximity to my inner mouth, out of which come my thoughts.

In reality this inner-man talk provides no help toward causal explanations. For the would-be cause (inner picture of such and such a character under such and such conditions of observation, etc.; or inner speech of this or that quality under these or those acoustical conditions) is not identifiable apart from the would-be effect (successful visualization, or knowledge of what one is thinking). And unless causes are identifiable apart from their alleged effects we have no basis for a causal explanation.[11]

---

[9] Augustine identifies the inner man indifferently as a man's soul (*anima*), his rational soul (*animus*), and his mind (*mens*). The quotation above contains an example of the first identification. Examples of the second and third can be found at *Confessiones* 10.6.9 and *In joannis evangelium* 99.4 respectively.

[10] See, e.g., *De trinitate* 11.3.6.

[11] One must be careful here. Peter Geach suggested to me, for purposes of comparison, the causal statement, "Smith craves to take cocaine because of taking cocaine." The concept *taking cocaine* is logically linked to the concept *craving to take cocaine*. And for this reason identifying the cause and identifying the effect cannot be completely independent of one another. Yet there is no conceptual reason to suppose that, on a specific occasion, we shall be unable to identify an instance of the cause, Smith's taking cocaine, as distinct from and independent of the effect, Smith's craving to take cocaine. ("Funny thing," Smith might honestly say, "I took my usual cocaine this morning and now I simply don't want any more of the stuff.") So my point is not just that the concept of a mental picture is logically linked to the concept of visualization; it is that on a given occasion the "presence" of a certain mental picture under good conditions of "mental observation" is not identifiable apart from the "resultant" visualization itself.

If this problem about identifying the would-be cause as distinct from the alleged effect is common to inner-man analogies generally (and I believe it is, though I shall not try to establish that here), we cannot say that inner-man analogies offer a basis for explaining our characteristically psychological abilities. And therefore we cannot say that the idioms of our language that express inner-man analogies commit us to dualism by laying bare the causes of our abilities to behave as we do.

Closely connected with the worry about how it is we know what we are thinking and what we intend to do is another line of argument worth mentioning here. Since a man can know what his inner voice is saying without knowing anything about his physiological states, since he can know what he says in his heart without even knowing what the physiological function of the heart is, it must be that a man who is reporting his thoughts and intentions is reporting on something completely non-physiological. Augustine presents the materials for this line of argument in this passage:

> For although we are said to think in the heart and we may know what we think without anyone else's knowing it, yet we do not know in what part of our body we have that heart where we (are said to) think, unless we learn it from someone else, who does not know what we are thinking.[12]

One thing to concern oneself with here is precisely the respects in which a report on a man's "outer" speaking can be properly considered a report on something physiological, or even something physical. (There is, for example, no way to record in ordinary German script the use of a uvular, as opposed to a lingual 'r', though they differ markedly in sound and certainly in the physiological mechanism used to produce them.) It may be that the inner-outer dichotomy encourages a rather oversimplified contrast between the physical (or physiological) and the non-physical (or non-physiological).

Anyway I shall not discuss this line of argument, not because it is unimportant, but because it rests fundamentally upon the very difficult matter of understanding avowals of one's own thinking, one's own intentions, etc.—whether or not these avowals are couched in the idioms of the inner man.[13] If they can be understood satisfactorily when purged of inner-man locutions, then surely to re-express them in terms of the inner man will not commit anyone to dualism.

I think the most interesting way in which inner-man locutions might commit us to dualism is through committing us to the contingency of the connection between the mind, its states and actions, and the body, and its states and actions. That is, if the contrast we draw through inner-man locutions between what a man does inwardly and what he does outwardly is such as to entail that the relation between mind and body is purely contingent, then inner-man locutions do entail dualism.

Before trying to decide whether or not they do, I should like to follow now in some detail what Augustine understands to be the nature of one important inner-outer contrast, viz., that between inner and outer speaking.

III

Thinking is inner speaking. Augustine mentions at least three different considerations that make this analogy appropriate. One thing is that we can, and often do, entertain verbal imagery. "The words of all audible languages," he reminds us, "are also thought in silence." The "rhythm of syllables" and the "tunes of songs . . . are present by means of their incorporeal images to those who think and turn over such things in silence" (*De trinitate* 15.11.20).

A second consideration is that we can and do use locutions like "said in his heart" and "said to himself" to report a man's thinking. Commenting on the Biblical locutions "they said, thinking" and "they thought, saying" Augustine says that both expressions mean "that they spoke within themselves, and in their heart, i.e., spoke by thinking" (*De trinitate* 15.10.17).

The third consideration arises from a certain picture of the mind or soul as something that gets the body to perform certain functions upon command. Thus "no one does anything willingly that he has not first said in his heart" (*De trinitate* 9.7.12).

---

[12] *De anima et ejus origine* (4.6.7).

[13] Thus Augustine presents a similar line of argument stripped of reference to the inner man in this passage: ". . . when the mind knows itself, it knows its own substance; and when it is certain about itself, it is certain about its own substance . . . Yet it is not at all certain whether it is air, or fire, or some body, or something bodily. Therefore it is not any of these." (*De trinitate* 10.10.16). Cf. Descartes, *Meditation* II.

I say Augustine mentions these three different considerations that make plausible the notion that thinking is inner speaking. But when he gets down to the business of telling us exactly what a true inner word is, it becomes clear that none of these three considerations makes plausible the notion that thinking is the utterance of such words as these. The true inner word of thought, he says, "is neither Greek nor Latin, nor of any other tongue." Ordinary "outer" words, even when they are entertained in verbal imagery, are but the "garment" of a thought. The true inner word is "couched in the heart, nude to the thinker, till it is clothed by the utterance of the speaker that it may proceed forth" (*Sermo* 187.3.3). Metaphysics imposes modesty; the word of thought cannot escape unclothed; but the clothing should never be confused with what it clothes.

Chief among the several reasons Augustine offers for insisting that the true inner word of thought is quite unlike any outer word, even one entertained in imagery, are these:

(1) One may have a thought before one has decided how to express it in "outer" words.[14]

(2) The same thought may be expressed by different "outer" words.[15]

(3) One may be able to pronounce "outer" words one does not understand, make slips of the tongue, babble, recite words by rote, etc.[16]

Suppose we concentrate on (3). Thinking, Augustine supposes, "underlies" meaningful speaking.[17] The speaking may, of course, go on without the thinking, or the speaking and the thinking may just fail to correspond. In the former case a man may be babbling or saying something by rote; in the latter case he may be dissembling, or using an expression he fails to understand, or making a slip of the tongue. Now if we suppose that thinking is inner speaking, we invite further regress. For the inner speaking, to be meaningful, would need to be backed by inner-inner speaking, etc., indefinitely.[18]

To stop the regress Augustine must say that there cannot possibly be such a thing as inner mouthing of inner words, inner speaking without comprehension, inner slips of the tongue, etc. (Hence the denial that a true inner word of thought can be identified with any verbal image; for a verbal image may be entertained without comprehension, etc.)[19] But with such possibilities ruled out completely, inner speech becomes unintelligible *as speech*. For a word is precisely the sort of thing that may be mouthed, pronounced mechanically, mistaken for another word, etc.

The notion that the true word of thought is itself outwardly unpronounceable undermines the three considerations listed above that make initially attractive the notion that thinking is inner speaking. First, the fact that we can and do entertain verbal imagery is no longer relevant. For a word of thought is not an ordinary word, or set of words, entertained in imagery.

Secondly, the fact that we can and do report a man's thinking as what he says to himself, in his heart, is no longer relevant either. For it is not as though what we so report could ever be a direct quotation. And to think of it as a translation, much as the Soviet Ambassador's remarks in Russian might be reported in English, perhaps within quotation marks, will not do either. For at least the Soviet newspapers may give the Ambassador's remarks in the original. But it is inconceivable that anyone should ever report a word of thought "in the original."

Thirdly, the picture of the mind or soul as something that gets the body to perform certain actions upon command becomes unintelligible once we realize that the mind is supposed to speak a metaphysically private mumbo-jumbo.

What Augustine is on to with his inner-speech analogy is the intimate relation between speaking and thinking; but he misconceives it. He wants us to view inner speaking (thinking) and outer speaking as species of a common genus—one an activity of the mind or inner man, and the other a corresponding activity of the body or outer man. But this will not do.

---

[14] See, e.g., *In joannis evangelium* 14.7.

[15] See, e.g., *Sermo* 288.3.

[16] See, e.g., *De trinitate* 13.20.26.

[17] Cf. Wittgenstein, *The Blue Book* (Oxford, Blackwell, 1958), pp. 34 ff.

[18] Cf. Gilbert Ryle, *The Concept of Mind* (London, Hutchinson, 1949), p. 296.

[19] When I speak here of stopping a further regress I am going beyond what Augustine actually says. What Augustine says is, e.g., that one who memorizes a string of Greek words without knowing what they mean remembers only what pertains to the outer man, i.e., he does not retain the thought that is expressed by these words (*De trinitate* 13.20.26). I mean for my talk of regress to be a way of stating in summary fashion Augustine's concern with the various differences there are between a thought and its expression and his insistence that "inner expressions" are, in these respects, like thoughts rather than expressions.

A man's thinking is what is expressed by what the man says when he says what is on his mind, or else by what the man *would* say if he *were* to say what is on his mind. The reference to speech in this formula (to either what a man says or what he would say if he were to speak his mind) may appear inessential. For the formula seems to come to this: a man's thinking is what is expressed by what expresses or would express his thinking; that is, a man's thinking is a man's thinking.

But the reference to speech is not eliminable in this way. Suppose a man says, ". . ." If he is awake, attending to business, has a command of the language, has not made a slip of the tongue, is sincere and candid, then these words express his thinking. It is not that when he succeeds in speaking his mind what he and we know to be expressed by these words matches the thought inside him. Or better, for what is expressed by his words to "match" the thinking "inside him" is just, in this simple case, for him to say what he says sincerely, with comprehension, etc.

A central role of the inner-speech metaphor is to help us relate a variety of situations to certain model cases. We relate situations in which one keeps one's thoughts to oneself, cases in which one is deceitful, cases in which one misspeaks, etc., to the model case of completely candid and successful speech. The inner-speech metaphor also helps us to relate to one another speeches in two different idioms (or in two different languages) as expressing the same thought. For the claim that somebody is thinking such and such can be understood as the claim that it is in certain respects *as if* he had said just that (sincerely, with comprehension, etc.), though perhaps actually he said something quite different or indeed nothing at all.

Thinking is not really a *kind* of speaking. But it is, if you like, *as-it-were* speaking.[20] That is, when a man is said to be thinking such and such, it is in certain respects (roughly specifiable in the context) *as if* he were saying just that.

IV

What I said in the last section about the functions of the inner-speech analogy will apply quite generally to a whole class of inner-outer contrasts. The class is made up of contrasts in those activities whose outer performance may be said, on occasion, to *express* a corresponding inner performance. A simple example is the inner smile. A man who is amused, say, by the lisps of a child and bites his lip to hide his amusement, may be said to smile inwardly. A mechanical smile, on the other hand, may be described as an outer smile with no corresponding inner smile. It is the man who smiles naturally and spontaneously who backs his outer smile with a corresponding inner one. And in his case the outer smile may be said to express his amusement (i.e., the inner smile).

This is a perfectly general approach to certain mental functions and psychological episodes. In the model case (the sincere speech uttered with comprehension, etc., or the natural and spontaneous smile) we speak of an inner act or activity corresponding to (and therefore expressed by) an outer act or activity. The locutions of inner-outer contrast are then a way of relating other situations to the model case.

It may be helpful to remind philosophers, preoccupied as they are with the mind-body problem, that the use of an inner-outer contrast sometimes has very little to do with whether there is anything mental "underlying" the physical. Augustine's contrast, mentioned earlier, between giving alms outwardly and giving alms inwardly is a case in point. Here the model case is the case of someone whose gift is a direct and uncalculating response to need. A penniless widow may be said to give inwardly, though she lacks the means to give outwardly. And the Pharisee, wanting to be admired, may give outwardly on a lavish scale without giving inwardly at all. Yet it is not that the Pharisee's giving is merely outward because it is mechanical and unthinking. The trouble is that it is wrongly motivated.

It is natural to ask whether the idioms of our language that express inner-man analogies could not be eliminated without loss to anyone but a rhetorician—whether the philosopher who concerns himself with expressions like "he said in his heart" and "he smiled inwardly" is not concerning himself with something that had best be left to philologists. A beginning reply would be "Yes, certainly they could be eliminated; 'he said to himself' can be replaced by 'he thought' and 'he smiled inwardly' by 'he was mildly amused without showing it.' "

---

[20] Cf. Peter Geach, *Mental Acts* (London, Routledge and Kegan Paul, 1957), p. 75, where "the concept *judging* is viewed as an analogical extension of the concept *saying*." See also Wilfred Sellars, "Empiricism and the Philosophy of Mind," *Minnesota Studies in the Philosophy of Science*, vol. 1, ed. by H. Feigl and M. Scriven (Minneapolis, University of Minnesota Press, 1956), pp. 253–329, especially pp. 307–311, 317–321.

Yet to handle the question this way is to miss the importance of these analogies. I am inclined to say that the concept of being amused is, in part anyway, an analogical extension of the notions of smiling and laughing and that the concept of thinking is, in part anyway,[21] an analogical extension of the concept of speaking. If I am right in saying this, then the fact that we can replace talk of inner smiles and inner speech with talk of amusement and thinking is rather uninteresting. The point is that our conception of what it is to be amused is built up from model cases of smiling, and our concept of thinking is built up from model cases of speaking. Being amused and thinking are, in this respect anyway, analogical concepts. And so also, in a similar way, are other concepts that are closely linked with inner-man locutions.

V

Augustine supposes that what we say about the inner man we say about an entity that animates the body for a time and leaves it at death. If he is right, then our very use of inner-man locutions such as "said in the heart" and "smiled inwardly" commits us to mind-body (or soul-body) dualism.

In Section III we followed Augustine's attempt to take seriously the dualism he believes to be inherent in one inner-man locution, "said in the heart," and its equivalents. We said that Augustine takes inner speaking to be something that animates outer speaking in such a way that when a stretch of inner speaking "underlies" a properly corresponding stretch of outer speaking, the result is that a person may be said to understand what he is saying, to be sincere in saying this, etc.—otherwise not. Augustine's development of the notion of inner speaking is a failure because inner speaking either itself needs special "backing" to be sincere, uttered with comprehension, etc., or else it must be thought to be automatically sincere, uttered with automatic comprehension, etc.; but in the latter case it is unintelligible as speech.

We could generalize our conclusion by saying that any attempt to understand inner *x*-ing as psychical "backing" for a purely physical operation (outer *x*-ing) will either invite further regress or else make inner *x*-ing unintelligible as *x*-ing.

In Sections III and IV I suggested an alternative approach to inner-man locutions. I suggested that they be interpreted as calling our attention to certain model cases of human behavior in such a way that when one is said to do something inwardly, it is in certain respects as if one had done this in conformity to the model case. And when one is said to do something merely outwardly, one's action fails to fit the model case in important respects, though it resembles it in a superficial way.

So the inner-outer contrast cannot be understood satisfactorily in the dualistic way Augustine would have us understand it. And it can be understood satisfactorily in quite another way. The question now remains, why should we have thought originally—not only that inner-man locutions might be *open* to dualistic interpretation—but that using them perhaps *commits* us to Augustinian dualism?

I think the answer is very simple. The contrast we draw between what a man does inwardly and what he does outwardly is such that there is a merely contingent relation between the inner and the outer man. That is, there will be, in general, no contradiction in ascribing one state or action to the inner man and a completely different, even opposite, state or action to the corresponding outer man. The inner man may smile while the outer man frowns or weeps. A person may be inwardly terrified, though outwardly calm and poised. The outer man may say, "What a charming evening!" while the inner man says, "A dreadful bore."

The contingency of this relation between what one may be said to do inwardly and what one may be said to do outwardly insinuates mind-body dualism. But we must be wary. Inner-outer contingency entails Augustinian mind-body independence only if we can identify the inner man with what Augustine takes to be the mind and the outer man with what he takes to be the body.

Let us remind ourselves what Augustine takes a mind to be (in this life, anyway). It is an entity whose activities underlie the being and behavior of the body in such a way as to make the difference between merely physical activity and the conscious, animated, meaningful, purposive behavior characteristic of living human beings. But what we have said about the locutions of inner *x*-ing means that the attempt to interpret them as reporting on the activities of what Augustine takes to be the mind is a failure. Yet the locutions have a perfectly good use in our language. And that use can be given a satisfactory alternative analysis. We must conclude that what Augustine takes to be a mind cannot be identified with something we are thought to report on when we speak of inner *x*-ing.

---

[21] I think there is more to these concepts than this. But I cannot hope to unravel the other strands here.

Nor can we identify what Augustine takes to be the body with the subject of the locutions of outer *x*-ing. Let us remind ourselves what an Augustinian body is. It is the "part" of a man that is buried at death. It has physical characteristics and may be moved in various ways. But no psychological predicates may be applied to it, or to its behavior.

Consider now an inner-outer contrast. Suppose we report that the outer man says, "What a charming evening!" while the inner man says, "A dreadful bore." What we are reporting is a case of dissembling. A person says one thing and thinks something quite different. But it is not true that the outer man's speech on this occasion is merely an unconscious physical movement, something completely "unbacked" by anything mental or psychical. What we have here is a case of dissembling. And the outer words, though insincere, are chosen, understood, and used quite purposefully. Otherwise they could not be said to be insincere. And so the subject of the outer speech is not an Augustinian body, but a dissembling person.

Take another example. A mother who is terrified of electrical storms is eager to prevent her children from developing the same phobia. When the children call her to their bedside during a storm she reassures them. She is outwardly calm, relaxed, and confident. But inwardly she quakes. Who or what is the subject of this outer calm? Not an unthinking physical object, but a mother who is really frightened but determined to behave in such a way as to hide her fear. Her outward calm is not a physical shell that conceals her feelings; it is a deliberate and self-conscious pattern of behavior that is meant to reassure her children.

One might have expected that "inner man," "outer man," and the whole family of inner-outer locutions would be technical terms in Augustine and that he would therefore not be bound to use them in conformity with our ordinary ways of talking. But it is quite clear, as I suggested at the beginning, that Augustine draws these locutions from the Bible, where they are certainly not used as technical terms. It is equally clear that Augustine means what he has to say about the inner man to be, among other things, an interpretation of these Biblical locutions. Moreover, the question before us is whether inner-man locutions, as we all use them (no doubt partly through the influence of the Bible), commit us to Augustine's dualism. And the answer to that question is clearly *No*.

University of Minnesota

Received January 10, 1966

American Philosophical Quarterly
Volume 4, Number 2, April 1967

X. TESTS FOR INTENTIONALITY

D. J. O'CONNOR

IT has been held by many philosophers that mental activities are basically different from physical activities. These philosophers have argued their case in various ways. One line of argument has been to construct tests which purport to distinguish ways of talking about mental activities and events from ways of talking about physical activities and events. A second approach is negative: Criticisms are offered of analyses which purport to show that mental events can be reduced to or explained in physical terms. We may call these the positive and negative cases for the thesis that mental activities are a well defined species of occurrences and that they cannot be accounted for in physical terms. I want to consider here some well-known examples of the positive case. They are taken from the writings of Professor Roderick M. Chisholm, who has expended more time and ingenuity on this problem than any other contemporary philosopher. I think that we may be confident that if Chisholm's case for the fundamental distinction between mental and physical is inadequate, no satisfactory case has yet been made.

Chisholm develops[1] a doctrine of the Austrian philosopher Franz Brentano. According to this doctrine, it is a feature peculiar to mental phenomena that they all contain in themselves an object or content. When we are conscious, we are conscious *of something*; when we believe, we believe *something*; and so on, for love, hate, desire, and the rest of what may loosely be called "mental states." This "intentional inexistence," as Brentano's translators call it, is found in all mental states and in no physical states. It therefore provides a defining characteristic whereby we may distinguish the mental from the physical. The trouble with this defining characteristic is, of course, that it is not always easy and obvious to say whether or not it is present in any particular phenomenon. Chisholm has therefore worked out certain tests for its presence. The tests are linguistic. There are certain properties peculiar to the sentences in which we talk about psychological states and events. In using these sentences we are displaying what Chisholm calls "the intentional use of language." This intentional use of language is never necessary when we talk about states and events which are not psychological. I shall first of all describe these tests and then make some critical comments about them. In doing so, I shall confine myself to one particular type of mental state, that of believing. It is possible that the Brentano-Chisholm doctrine may be more defensible for some mental states than for others; but if I can show that it does not hold for belief, I shall at least have shown that it does not provide an infallible indicator for distinguishing the mental from the physical.

Chisholm claims that a simple declarative sentence is intentional if it satisfies any one of the following three criteria:

(1) Suppose that a sentence S uses a substantival expression, that is, a noun or a description, in such a way that neither S nor its negation implies that there exists or that there does not exist anything to which the substantival expression applies. For example:

Diogenes looked for an honest man

does not imply either that there are or that there are not any honest men, nor does its contradictory:

Diogenes did not look for an honest man.

Such sentences are intentional. By contrast, the sentence:

Diogenes lived in a tub

is not intentional, for it does imply the existence of a tub.

(2) A second criterion, equivalent for many examples to the first, but with a slightly different range of application, is the following. If a sentence S contains a propositional clause and neither S nor its contradictory implies that the propositional clause is true or false, then S is intentional. For example, neither the sentence:

I believe that there is a monster in Loch Ness

nor its contradictory:

---

[1] *Perceiving: A Philosophical Study* (Ithaca, 1957), Ch. 11 and "Sentences about Believing" in *Proceedings of the Aristotelian Society*, vol. 56 (1955–56), pp. 125–148.

I do not believe that there is a monster in Loch Ness

implies that there is a monster in Loch Ness or that there is not.

(3) Finally, let us suppose that we can describe something by alternative names or descriptions. For example:

Aconcagua is the highest mountain in the Andes.
The morning star is the evening star.
The author of *Marmion* was the author of *Waverley*.

Let us call such a sentence *E* and let us suppose that *A* is a sentence using one of these names or descriptions and that *B* is a sentence like *A* except that it uses the alternative description. Then if the conjunction of *A* and *E* does not imply *B*, *A* is intentional. For example:

George IV believed that Scott was the author of *Marmion*

together with:

The author of *Marmion* was the author of *Waverley*

does not imply:

George IV believed that Scott was the author of *Waverley*.

This makes the first sentence intentional, although the sentence

Scott was the author of *Marmion*

together with the equivalent sentence *does* imply

Scott was the author of *Waverley*.

Thus the sentence:

Scott was the author of *Marmion*

is not intentional by this criterion, or indeed by any other.

Chisholm points out that these three criteria and, in particular, the first two, "overlap to a considerable extent" in their application. Any one of them would be regarded by Chisholm as a sufficient condition for intentionality. I think he would agree, however, that none of them is a necessary condition. (And this, as will be seen later, is important.) For it is easy to produce examples of sentences that are intentional by one criterion and not intentional by another. For example:

I believe that there is a monster in Loch Ness

is intentional by criterion (2). But, paradoxically:

I know that there is a monster in Loch Ness

is not intentional by this criterion, since on any usual analysis "I know that *p*" does imply that *p* is true. (And by "imply" here I mean that the proposition "*p* is true" is one of the propositions whose conjunction forms the analysis.) However, in this case at least, we can save ourselves from the paradox of maintaining that "I know that *p*" is not an intentional use of language by appealing to the third criterion. Let us suppose that there is a monster in Loch Ness and that it is, in fact, the last specimen of *Ichthyosaurus communis* surviving somehow from the Jurassic age. And suppose that I have often seen this creature but know nothing of its zoological affinities. Then we might have:

(A) I know that there is a monster in Loch Ness.
(E) The monster in Loch Ness is the last surviving specimen of *I. communis*.

But (A) and (E) together would not imply (B):

(B) I know that the last surviving specimen of *I. communis* is in Loch Ness.

Thus (A) would be intentional by criterion (3). There would still be sentences using the word "know" which do not seem to be intentional by any of Chisholm's criteria. (For example: "I know that cats like fish.") But I shall not pursue this here as I wish to confine the discussion, for simplicity, to the single allegedly intentional concept—belief.

I want first to make some general remarks about the nature of the argument that Chisholm is using here. He is trying to show that, in the English language at least, there are certain modes of expression that are appropriate to what are ordinarily regarded as psychological events and activities and inappropriate to *prima facie* physical events and processes and *vice versa*. Now suppose that this thesis were established. What would it prove? In particular, could it be regarded as evidence that the distinction between mental and physical was any better grounded than it was before the intentional use of language had been observed? Surely not. The most that we could say would be that the structure of language does reflect our unsophisticated beliefs on this point. But it would have no tendency whatever to show that the distinction between mental and physical was in any way basic or that the mental could not be reduced to or explained in terms of the physical. So even if the distinction could be shown to be well founded, it would be no more than an interesting fact about language. And this would still be the case, if it could be shown that an analogous distinction could be drawn in most other natural languages. (Of course, nothing of the kind has been established: so that, for all we know, this is a peculiarity of English and a few other closely related languages.) But however widespread the distinction between intentional and nonintentional language could be shown to be, it could have no interesting philosophical consequences unless it could be proved that it was not merely a prescientific or false scientific view of the world fossilized in conventional usage. (For example, "the sun rises" dates from pre-Copernican astronomy, "oxygen" from a superseded chemical theory, and so on.) Perhaps our language merely embodies a discredited Cartesianism. Professor Chisholm does not in fact make any strong metaphysical claims that depend upon his theory of intentionality. And he would, of course, want to say that the intentional use of language is very different, in one important way, from the fossilized errors that I have referred to. We can easily redescribe the sun's rising and similar misdescribed phenomena in correct descriptive language if it becomes worth our while to do so. But it is, according to Chisholm, necessary to use intentional language to describe mental phenomena and activities unless we introduce special terms which we do not need for the description of physical phenomena. Conversely, it is never necessary to use intentional locutions or terms when we talk about physical events and processes. We can always translate such sentences into nonintentional language without loss of meaning. Whether or not such claims are true is a question quite separate from Chisholm's account of intentional language. Only if this survives criticism would it be worthwhile taking this point further.

I shall look, first of all, at his first and second criteria. They may, as he says himself, be regarded as substantially equivalent for most purposes. It is, in fact, the second which applies most naturally to belief-sentences. Any belief-sentences that are framed so as to meet the first criterion can be restated in an equivalent form to meet the second. (For example, "I believe in ghosts" can be rephrased as "I believe that ghosts exist." Chisholm's criterion for this type of sentence is that it is intentional if neither the sentence nor its contradictory implies that the propositional clause is true or that it is false. For example:

I believe that there is a monster in Loch Ness

does not imply that the sentence:

There is a monster in Loch Ness

is true or that it is false. And nor does the sentence:

I do not believe that there is a monster in Loch Ness.

Using *P* for the belief sentence and *Q* for the propositional clause, Chisholm claims that it is false that *P* implies *Q*, false that not-*P* implies *Q*, false that *P* implies not-*Q* and false also that not-*P* implies not-*Q*.

It is not possible to know what this means unless we know what sense he is putting on the word "implies." It is obvious at least that he cannot be using the word in the sense of "materially implies" familiar from propositional logic. For if he was, he would be affirming a logical contradiction. The compound proposition "it is false that *P* implies *Q* and false that not-*P* implies *Q*" can easily be shown to be equivalent to the proposition:

It is false that *P* and not-*P* together imply *Q*.

And this last proposition is logically false. For it is a truth of propositional logic that the contradictory conjunction "*P* and not-*P*" materially implies any proposition whatever. So clearly Chisholm intends the word "implies" in this context to be taken in some stronger sense. But what sense?

We may dismiss, first of all, the so-called "informal" senses of implication which are exemplified in the conventional empty examples like " 'John is a bachelor' implies 'John is not married' " or " 'Red is a color' implies 'Red is extended'." In so far as such alleged implications are genuine, they are instantations of well-known logical truths "*P* and *Q* imply *P*" and the like. In all such cases, the implication is justified by a logical apparatus in the background which is taken for granted. ("Informal logic," if it is to be reputable, cannot be other than formal logic with steps skipped for economy. It is not another species of logic for the use of those who wish to be able to argue rigorously without taking trouble.)

Let us assume then a stronger sense of "implies" than the material implication of truth-functional logic. C. I. Lewis' "strict implication" is an obvious alternative. Then we have:

It is false that *P* strictly implies *Q*

which a little manipulation shows to be equivalent to:

The conjunction of *P* and not-*Q* is possible.

Likewise:

It is false that not-*P* strictly implies *Q*

becomes:

The conjunction of not-*P* and not-*Q* is possible.

And we may analyze the other two theses of his criterion in the same way to get:

The conjunction of *P* and *Q* is possible

and:

The conjunction of not-*P* and *Q* is possible.

(That this is a plausible interpretation can be seen from the fact that if we say that the sentence

"Diogenes lived in a tub" implies that there was a tub, we do mean that it was impossible that Diogenes lived in a tub and that no tub existed.)

But although this interpretation of Chisholm's "implies" as Lewis' relation of strict implication seems at first sight to offer an acceptable account of the theses about intentionality under consideration, it is still open to objection. There are two possible interpretations of the phrases "The conjunction of P and not-Q is possible" and its variants. The first is to understand such statements simply at their face value as affirming the possibility of two propositions being simultaneously true. The second is to understand such affirmations as meaning that P and Q are *logically independent* of each other so that the truth value of P or its negation is entirely irrelevant to the truth value of Q or its negation, and vice versa.

The first interpretation may easily be refuted by constructing the following counter-example. It is possible to believe that logically false propositions are in fact true. For example, Thomas Hobbes believed that it was possible to square the circle. Then if P is the proposition that Mr. X believes that the circle can be squared and Q is the proposition "The circle can be squared" then one of the tests proposed by Chisholm is the following:

It is false that P strictly implies not-Q.

And this is equivalent to:

It is possible that the conjunction of P and Q is true.

But since Q stands for the logically false proposition

The circle can be squared

the conjunction can never be true. Thus "implies" cannot mean "strictly implies" in the context of Chisholm's second criterion, provided at least that we interpret "P strictly implies Q" as usual to mean "It is not possible that P is true and Q false."

But what of the second interpretation, that P and Q are logically independent? If by "logically independent" we mean that the truth-value of P or its negation has no bearing at all on the truth value of Q or its negation where P is a proposition of the form "X believes Q" and conversely, the truth-value of Q has no bearing on that of P, the answer is that it is an empirical matter that such logical independence can never in fact hold. Although it may be true that we are never able to say, of any particular instantiations of P and Q what bearings they have on each other's truth-values, it is nevertheless quite certain that if our beliefs were not very often true (and indeed more often true than not) we should in fact cease to hold beliefs at all. It is the ascertained truths of some of our beliefs that reinforces our tendency to hold others. Thus "I believe Q" can be said to afford inductive evidence for the truth of Q, though the probability of Q, on the evidence that I believe it, is something that would have to be estimated separately in every case. Thus to say that P and Q are logically independent where P is a belief sentence and Q its object is to make a claim that cannot be supported. P is always evidence (of an inductive kind) for Q, though of course it is never conclusive evidence. And this means that the truth or falsity of P does have *some* bearing on the truth or falsity of Q. And if this is so, in what sense can they properly be said to be logically independent? We can perhaps say of any given collection of belief statements and their respective internal accusatives that *distributively* they appear to be independent of one another in that the truth or falsity of any particular P_k gives us no clue to the truth of its corresponding Q_k. But *collectively*, the truth of the objects of our beliefs is not independent of the truth of the statement that we believe them. The truth of the belief statement confers a certain probability on the truth of its object and vice versa. And this in turn erodes the purity of the relation of logical independence that is alleged to hold between them. For even if the probability of Q, given the truth of P, is very small, say 0.001, we can hardly claim that the truth-values of belief statement and its object are independent. One does have a bearing on the other.

Let us suppose finally that:

P implies Q

means here

Q is derivable from P.

Chisholm's second criterion then becomes: If P is a belief-sentence such as "I believe that there is a monster in Loch Ness" and Q the statement of the propositional clause, that is, "There is a monster in Loch Ness," it is false that Q is derivable from P or from not-P and likewise false that the negation of Q is derivable from P or from not-P.

To this, our reply must be: One proposition is derivable from another only within the context of some logical system. To claim that Q *is* derivable from P is to invite the challenge: Show me the derivation. Q is never derivable from P alone. (Even in the trivial case where P is itself a conjunction of Q and some other proposition, we need a rule of simplification or the like.) Therefore we need to know the apparatus of rules and axioms that is being assumed and to have the derivation performed. But to claim that *Q* is *not* derivable from *P* is to make a much larger claim and one that is correspondingly more difficult to substantiate. How can we show that a proposition *Q* is not derivable from *P* within a certain system *S*? First, we might try to show that *Q* is a well-formed expression of *S* but that the available decision procedures show that *Q* is not a derivable proposition. And this method would be available only if we had a decision procedure for the level of logic in question. And since belief-sentences cannot be formulated within the decidable levels of logic, namely propositional logic and the simpler parts of predicate logic, such a demonstration of the non-derivation of *Q* from *P* would not be available. Secondly, it might be that a defender of intentional language might appeal to this fact, the fact namely that belief-sentences cannot be formulated within classical predicate logic as a proof that *Q* is not derivable from *P*. But it is a trivial (and easily remediable) form of non-derivability that stems from the fact that the sentence in question cannot be formulated at a particular level of logic. No philosophical consequences follow from this any more than they do from the fact that a syllogism cannot be formulated in the language of the propositional calculus. All we have to do is to extend our logic to incorporate sentences of the form "*X* believes *P*." We will then be in a position to attend seriously to the question: Is *Q* derivable from *P*? (The development of such intentional formal logics is still in a primitive stage. But attempts at such formulations have been made and it would be imprudent to anticipate the future.) We must remember, however, that even then it can be a very difficult task to show that a particular sentence which can be formulated at a particular level of logic cannot be proved at that level. (Remember Gödel's theorem.) It is one thing to fail to prove a proposition and quite another to show that it is in fact unprovable. We may take it then that derivability does not offer us the explication of Chisholm's term "implies" that we are seeking.

My criticism of Chisholm's second criterion is therefore as follows. It is too vague to be of use as long as we are given no clear sense in which the word "implies" is to be taken. And if we try to assign plausible senses to the term, his claim turns out to be either unproven (and perhaps unprovable) or else simply false.

Let us now look more briefly at the other criterion for the intentional use of language. By this criterion, the sentence:

> George IV believed that Scott was the author of *Marmion*

is intentional because taken together with the sentence:

> The author of *Marmion* was the author of *Waverley*,

it does not imply:

> George IV believed that Scott was the author of *Waverley*.

In other words, although sentences of the form "*A* is *B*" and "*B* is identical with *C*" together imply "*A* is *C*," sentences of the form "*X* believes that *A* is *B*" together with "*B* is identical with *C*" do not imply a conclusion of the form "*X* believes that *A* is *C*." Can we take this as a sufficient condition of intentionality for the sentence that forms the first premiss?

We may note at once that the same questions about the meaning of "implies" must be raised here as were raised previously. And unless we can clarify the meaning, the criterion will be useless. Now if "it is false that *P* implies *Q*" is to be read "*Q* is not deducible from *P*" we shall want to know in what logical system the impossibility of deducing *Q* from *P* is to be demonstrated. And then, we can reasonably ask for the demonstration to be performed. Some of the difficulties in the way of making good this claim for non-derivability have already been examined.

Suppose, however, we interpret "implies" to mean "strictly implies." This interpretation fares better when applied to the third criterion than it does when applied to the second. It does, however, lead to some paradoxical consequences. Any belief-sentences tested by this criterion reduce to the following statement: It is possible that three sentences of the following form are all true:

> *X* believes that *A* is *B*
> *B* is identical with *C*
> *A* does not believe that *A* is *C*.

And this of course does sometimes happen, as in the case of George IV's beliefs about the authorship of *Waverley* and *Marmion*, cited above. But equally it often happens that *X* believes that *A* is *B*, *B* is identical with *C*, and *X* believes that *A* is *C*. For example, the man who was allied commander in 1944 was president of the United States in 1954; I believe that Eisenhower was allied commander in 1944 and I also believe (after the event) that Eisenhower was president in 1954.

Thus compound propositions of the following type are often true:

> It is possible that X believes that A is B, that B is identical with C and that X believes that A is C.

And this is equivalent to the proposition:

> It is false that X believes that A is B and that B is identical with C together strictly imply that *it is false that* X believes that A is C.

And this does not satisfy Chisholm's third criterion. So that by this criterion, whether a sentence is intentional or not will depend on whether the person who believes P also believes Q, where P differs from Q only by the substitution of one synonymous expression for another.

But it is surely paradoxical to suppose that belief sentences (and other intentional sentences about knowing and remembering) lose their intentional status with the acquisition of greater knowledge. But perhaps it will be objected that it is wrong to say that a sentence can lose its intentional status in this way. For Chisholm's third criterion says only that if a sentence satisfies the criterion, then it is intentional and not that it is not intentional if it fails to satisfy it. A sentence cannot lose a status that it never had. Exactly so: we must remember that these tests are proposed as tests *for intentionality*. We are, supposedly, in the dark as to whether we may promote a sentence to the status of being intentional until we see if it satisfies one of these tests. And if it satisfies none of them, as in the example above, we have no grounds for claiming that it is intentional. For this, we need further independent tests which we do not have. Does it not at least seem paradoxical that the intentionality of a sentence, on this criterion, seems to turn on the question of *who* believes the propositional clause in question or *when* he believes it? If, for example, we read "Constable," the name of Scott's publisher, for "George IV" in the Scott-Waverley example, the belief sentence is not shown to be intentional by this criterion (though it is not shown not to be intentional either). (Are we to assume that "intentional" is a predicate which does not obey the law of excluded middle?)

Thus it seems that an examination of the working of Chisholm's criteria for distinguishing sentences which are intentional from those which are not shows that they do not perform the task for which they were designed. Perhaps they can be re-formulated to work more effectively. Or perhaps the distinction which they are supposed to mark does not, in fact, exist.

University of Exeter

Received April 7, 1966

AMERICAN PHILOSOPHICAL QUARTERLY

Edited by

NICHOLAS RESCHER

With the advice and assistance of the Board of Editorial Consultants:

William Alston
Alan R. Anderson
Kurt Baier
Lewis W. Beck
Nuel D. Belnap, Jr.
Roderick M. Chisholm
L. Jonathan Cohen
James Collins
James M. Edie
Peter Thomas Geach
Adolf Grünbaum
Carl G. Hempel
John Hospers
Raymond Klibansky
Ernan McMullin
Benson Mates
John A. Passmore
Günther Patzig
Richard H. Popkin
Wesley C. Salmon
George A. Schrader
Wilfrid Sellars
Alexander Sesonske
J. J. C. Smart
Manley H. Thompson, Jr.
James F. Thomson
G. H. von Wright
John W. Yolton

VOLUME 4/NUMBER 3

JULY 1967

CONTENTS

I. HARALD OFSTAD: *Recent Work on the Free-will Problem* 179

II. ROBERT C. COBURN: *Persons and Psychological Concepts* . . . 208

III. VIRGIL C. ALDRICH: *On Seeing Bodily Movements as Actions* . 222

IV. CHARLES B. DANIELS: *Colors and Sensations: Or How to Define a Pain Ostensively* 231

V. JOHN L. POLLOCK: *Mathematical Proof* 238

VI. J. W. SWANSON: *Religious Discourse and Rational Preference Rankings* . 245

VII. GEORGE NAKHNIKIAN: *The Cartesian Circle Revisited* 251

PUBLISHED BY BASIL BLACKWELL WITH THE COOPERATION OF THE UNIVERSITY OF PITTSBURGH

AMERICAN PHILOSOPHICAL QUARTERLY

POLICY

The *American Philosophical Quarterly* welcomes articles by philosophers of any country on any aspect of philosophy, substantive or historical. However, only self-sufficient articles will be published, and not news items, book reviews, critical notices, or "discussion notes."

MANUSCRIPTS

Contributions may be as short as 2,000 words or as long as 25,000. All manuscripts should be typewritten with wide margins, and at least double spacing between the lines. Footnotes should be used sparingly and should be numbered consecutively. They should also be typed with wide margins and double spacing. The original copy, not a carbon, should be submitted; authors should always retain at least one copy of their articles.

COMMUNICATIONS

Articles for publication, and all other editorial communications and enquiries, should be addressed to: The Editor, *American Philosophical Quarterly*, Department of Philosophy, University of Pittsburgh, Pittsburgh, Pennsylvania 15213.

OFFPRINTS

Authors will receive gratis two copies of the issue containing their contribution. Offprints can be purchased at low cost through arrangements made when checking proof.

SUBSCRIPTIONS

The price *per annum* is six dollars for individual subscribers and ten dollars for institutions. Checks and money orders should be made payable to the *American Philosophical Quarterly*. Back issues are sold at the rate of two dollars to individuals, and three dollars to institutions. All correspondence regarding subscriptions and back orders should be addressed directly to the publisher (Basil Blackwell, Broad Street, Oxford, England).

American Philosophical Quarterly
Volume 4, Number 3, July 1967

I. RECENT WORK ON THE FREE-WILL PROBLEM

HARALD OFSTAD

I. Some Major Trends

ALTHOUGH the problems formulated and the answers given by contributors to the free-will discussion during the last decade[1] vary immensely, there are nevertheless certain noticeable trends. For instance, very few of the authors take it as *a priori* or empirically certain that the so-called principle of causality can be applied to human decisions and actions.[2] It is mostly considered as a working hypothesis.[3] Moreover, due to the obvious difficulties of discussing the question of causation as an ontological issue, the authors tend to concentrate on the question whether or not our decisions and actions can be predicted or explained. In this connection the Bergsonian point, that a chooser cannot predict his future decision without making it,[4] has been taken up again by several authors. Some of them also stress the methodological difference between things and human beings, denying the program of the unity of science movement, and advocating some variation of the old distinction between *verstehende* and *erklärende* explanations,[5] emphasizing the methodological dissimilarity between explanations of actions (in terms of reasons), and explanations of events (in terms of causes). Such concepts as reason, choice, decision, intention, and action have been subjected to numerous analyses, some of which indirectly may be of value to the free-will discussion.

At the same time that we find this increased scepticism toward certain interpretations of the so-called deterministic doctrine, we find that quite a few of the so-called indeterminists and libertarians accept that freedom in an indeterministic sense is not a sufficient, only a necessary condition of freedom in the sense required by moral responsibility. It is also necessary that the agent was able to do something, e.g., to make what has been called a certain "effort of will."

The period includes some able defenders of so-called mild (soft) as well as hard determinism, but they are, on the whole, less concerned about proving that our decisions and actions are caused than trying to show that moral responsibility is—or is not—compatible with this assumption. Moreover, also the determinists have become interested in the analysis of such expressions as "tried," "was able to," "had the ability to," "could have acted (decided) otherwise," "had it in one's power to act (decide) otherwise," rather than following their more traditional line only, defining "freedom" as absence of compulsion or constraint. Thereby they have come closer to the libertarian position in so far as this stresses "the ability to make efforts of will." This new-won interest, both of some determinists and indeterminists, in the analysis of the sense in which we have it in our power to decide and act otherwise than we do, and the extent to which we have such power, is one of the most promising features of the recent discussion. My main reason for this evaluation is that this line of thinking seems to me to focus on that sense of "free" which is most relevant from the point of view of moral responsibility. So even disregarding the semantic unclarity of such terms as "determinism" and "indeterminism," they have become less significant also because they do not refer us to the crucial issues.[6]

---

[1] This, of course, is quite an arbitrary delimitation of a period of debate.

[2] Blanshard is one of the few exceptions. See his "The Case for Determinism" in *Hook* (1958). (My references will include only the name of the author, and, if convenient, the year of the publication of his work. The exact reference can be found by consulting the bibliography. If one author has several relevant publications from the same year, they will be marked, e.g., 1960.1, 1960.2, etc.). I see no reason to define "the principle of causality" in an article of this kind. Different definitions are discussed in the literature referred to.

[3] See, e.g., *Nagel* (1961), pp. 316 ff.

[4] Bergson, *Time and Free Will* (London, 1910), pp. 184 ff.

[5] Cp. W. Dilthey, "Ideen über eine beschreibende und zergliedernde Psychologie," published 1894, reprinted in *Gesammelte Schriften* (Leipzig und Berlin, 1924), Band V, pp. 139–240, and H. Rickert, *Kulturwissenschaft und Naturwissenschaft* (Tübingen, 1915).

One of the great weaknesses of the recent discussion is the insufficient clarification of the moral context in relation to which the authors maintain that we are (or are not) free. Though quite a few are aware of the necessity of clarifying such expressions as "moral responsibility" in connection with the free-will dispute, the analysis is still at a preliminary level, and for some strange reason the overwhelming majority of writers[7] consider moral responsibility equivalent with moral blameworthiness—as if every philosophy of responsibility commits us to some version of a "you-naughty-boy"-philosophy. Obviously, we may hold a person morally responsible without holding him blameworthy, we may, e.g., consider him morally guilty without considering him blameworthy. Moreover, in spite of the awareness of the role of excuses, very few have seen the moral relativity of the free-will issue. Some philosophers even seem to believe that they can solve the so-called free-will problem by finding out how we use such sentences as "He could have decided (acted) otherwise." But even those who understand that such investigations are not decisive and that we must find out whether or not we are free in the sense required by moral responsibility are insufficiently aware of the fact that this sense may vary with the nature of the moral system within which the concept of responsibility is defined. According to the moral system E_1, a person is perhaps not responsible unless he decided freely in sense x, whereas, according to E_2, he is not responsible unless he decided freely in sense y. Let us suppose that we are free in sense x but not in sense y. Do we then have a free will? That depends upon whether the system E_1 is "more correct than" the system E_2. Hence, there is an important sense in which a solution of the free-will issue presupposes a clarification and solution of certain basic problems within ethics.

II. Freedom and Moral Responsibility

To simplify the comparisons of different positions, let us—whenever it is convenient—concentrate on an example of an action which we shall assume to be morally wrong: on May 1, 1965, Mr. Smith shot his aunt Olga to death. He admits whatever can be shown by a film of the happenings, but he maintains that he is not responsible because his will was not free. How would the recent contributors to the free-will debate deal with this contention?

Most of them would consider it impossible to take any standpoint to it until it has been clarified what Smith meant by saying "my will was not free." By the term "will" he may refer, among other things, to his so-called volitional capacity, his choice, his decision or action. Let us suppose that he meant that his action was not free. But what did he mean by "not free"? Perhaps he meant that his action was compelled in some way, or that it was caused, or caused by certain special factors, e.g., neuro-physiological events or the so-called irrational part of himself, or perhaps he meant that he did not have it in his power to act otherwise than he did.

Let us suppose that Smith meant that his action was caused (T). We can then raise the questions: (*i*) Is T sufficiently precise, and (*ii*) Is T true? But, as indicated above, a more fundamental question is: (*iii*) Is the lack of freedom in this sense of "freedom" sufficient to exclude moral responsibility? Or, formulated in more general terms: In what sense of "freedom" is absence of freedom a sufficient reason for holding that Smith is not morally responsible? The answer will depend upon our concept of responsibility and the way in which we justify statements ascribing responsibility to agents. It is impossible in this article to deal with these problems in any thorough way, but certain basic distinctions must be made.

When someone says that Smith is responsible for the death of Olga, he may mean only (*i*) that some of Smith's muscular movements were a causal factor in that series of events which led to her death. The statement that Smith is responsible in this sense is empirically true or false, and there is no plausible sense of "freedom" in which freedom is a condition of responsibility in this sense. Next, one may mean (*ii*) that her death was due to something Smith did, where "doing something" at least implies that Smith was sufficiently conscious to understand that his behavior was dangerous.[8] In relation to this sense it is appropriate to ask whether freedom in any sense is a condition of responsibility. Furthermore, one may mean (*iii*) that Smith is morally guilty. The concept of moral guilt is rather neglected in recent literature, and in so far as it is taken up at all, the authors tend to conceive of it as identical with blameworthiness. But this is not acceptable, for a person who is against moral blame as a matter of principle—and therefore considers nobody blameworthy—may nevertheless consider someone guilty. If he finds that the utterance of the sentence "You are guilty" will be experienced as moral blame, he may stop saying such things, but still think them.

---

[6] Cp., e.g., *Pears* (1963), ch. 4.

[7] Taylor is one of the exceptions. See his "Determinism and the Theory of Agency" in *Hook* (1958), p. 226.

[8] Somewhat similar distinctions are made by *Feinberg* (1965). See also, e.g., *Brandt* (1958), *Friedrich* (1960), and *Hospers* (1961). A useful discussion of the concept of criminal responsibility can be found in *Wootton* (1959).

To say that Smith is morally guilty of the death of Olga is to give a moral evaluation of Smith as the producer of this result in view of all the factors in favor and disfavor of him. And, analogously to criminal guilt, moral guilt can vary in degrees. Hence, saying that Smith is morally guilty of killing Olga can be expressed thus: "In killing Olga you were morally bad to such and such a degree." To analyze the sense, if any, in which freedom is a condition of moral guilt, is a crucial task within the free-will debate.

To say that Smith is morally responsible may also mean (*iv*) that he is morally blameworthy. By "blameworthy" one often seems to mean that the agent—in my terminology—is morally guilty, and that we consequently have a reason for blaming him. I have already mentioned the relation between freedom and guilt, and the only new question raised by (*iv*) is therefore concerned with the relation between freedom and blame. This problem, however, can be more adequately dealt with by turning to our fifth interpretation of saying that Smith is responsible: someone (but not necessarily *you* or *I*) ought to express moral disapproval of him (*v*). And the final question is this: In what sense, if any, of the term "freedom" is freedom a condition of moral blame?

My exposition must be limited to some of these senses of "responsible," and by concentrating on sense (*iii*)—moral guilt, and (*v*)—moral blame, we seem to have the best chance of getting at the most important issues. Our problems can now be formulated thus: How do the contributors to the recent free-will debate deal with the question whether freedom in any sense is a condition of moral guilt and/or moral blame? What do they mean by "moral guilt," "moral blame," and by "freedom"? How do they defend that freedom is—or is not—a condition of guilt or blame? And how do they defend that this condition is—or is not—fulfilled?

Connected with the debate concerning the correct analysis of "free," there is another concerning the object of freedom.[9] Some maintain that the important question is whether or not our actions can be free, others that the relevant issue is whether or not the self, the acting agent, or our decisions are free.

In view of the concepts used by the majority of the participants in the recent debate, it seems preferable to choose "decision" and "action" as our key terms. But if we take "freedom" in the sense of power to decide or act otherwise, talking about the freedom of decisions or actions is really to talk about the *agent's* power to decide or act.

III. Types of Positions

Before dealing with some representative contributors, it may be useful to present a short survey of the major types of positions.

1. *Indeterminism.* Let us distinguish two variations:

1.1. *Ontological indeterminism:* decisions are not caused, or, at least, some of their characteristics are not caused.

1.2. *Epistemic indeterminism:* it is impossible to predict (and explain) decisions, or, at least, some of their characteristics cannot be predicted (and explained), at least not by the chooser when he is deliberating.

2. *Two-domainism:* the causal principle does not apply to decisions or actions, they are neither caused nor uncaused, but it applies to bodily movements —they are either caused or uncaused.

3. *Libertarianism.* Let us distinguish two types:

3.1. *Moral libertarianism:* our decisions between following our interests or doing what we consider our duty are made by a self which is not completely causally determined, and we are always able (unless we are mentally sick) to put forth the effort required in order to decide in favor of the latter.

3.2. *Neutral libertarianism:* our decisions between different courses of actions are made by a self which is not completely causally determined, and we are always able (unless we are mentally sick) to put forth the effort required in order to decide in favor of anyone of them.

4. *Fatalism:* I distinguish between:

4.1. *Logical fatalism:* it is logically necessary that the future must be what it is going to be, and logically impossible that it can be anything other than what it is going to be.

---

[9] See, e.g., *Thalberg* (1964).

4.2. *Empirical fatalism:* our deliberations do not influence our decisions and actions.[10]

5. *Hard determinism:* moral responsibility presupposes that we, in a categorical sense of "could," could have decided or acted otherwise, but since our decisions and actions are caused, it follows that we are not free in this sense.

Two variations may be distinguished, depending upon the moral consequences drawn. According to the first:

5.1. *Hard determinism of the anti-moral-blame type:* the above view implies that we are never justified in *morally* blaming anyone, but we may be justified in making use of non-moral blame.

According to the second:

5.2. *Hard determinism of the anti-merit-blame type:* the above view implies that we are never justified in maintaining that anyone *deserves* or *merits* blame, but blame, moral as well as non-moral, may be justified, e.g., in terms of utility.

Some authors have defended the position that the moral consequences mentioned in 5.1 and 5.2 follow, not from the principle of causality, but from the principle that we are products of hereditary and environmental factors. We may refer to this position as:

6. *Hard evolutionism.*

7. *Mild determinism:* common to the positions referred to by this label is the contention that moral responsibility does not require that we are free in any sense which is incompatible with the assumption that our decisions and actions are caused. Some of the adherents to this view hold that responsibility presupposes that our decisions are caused, whereas others deny that it does.

Three variations may be distinguished, depending upon the sense of "freedom" considered crucial.

7.1. *Freedom as modifiability:* according to this position, the only requirement is that the agent is free in the sense that the use of blame will modify his behavior in the future. Knowledge about his past, e.g., about whether or not he could have decided otherwise, is of interest only in so far as it helps us to judge his present modifiability.

7.2. *Freedom as hypothetical behavior:* moral responsibility requires that the wrongdoer would have decided or acted otherwise if this or that condition had been fulfilled, e.g., if he had chosen otherwise. According to this view, we should, ideally speaking, know something about the agent's past.

7.3. *Freedom as categorical power:* according to this view, moral guilt, in order to be maximal, requires that the agent had it in his power to decide or act otherwise—in the fullest sense in which we can ever have anything in our power. And although our power to decide and act may be limited by many different factors, it is never the principle of causality which annihilates or delimits it.

IV. Ontological or Epistemic Indeterminism

The main arguments presented by recent authors accepting ontological or epistemic indeterminism are not based on empirical research or introspective reports about what we find ourselves believing when we choose, but on some kind of philosophical analysis of such concepts as intention, decision, and voluntary action. By the use of such analyses, they have tried to discover a logical proof for indeterminacy or unpredictability.

C. Ginet

In their article "Decision, Intention and Certainty" (1958), Hampshire and Hart tried to show, among other things, that an agent cannot both know what he is going to do and deliberate about what to do. A similar line of thinking was later presented by *Ginet* (1962). He wanted to prove that

(*i*) It is conceptually impossible that a decision should be caused.

Ginet maintains that (*i*) follows from the following two propositions:

(*ii*) If it were conceptually possible for a decision to be caused, then it would be conceptually possible for a person to know his own decision before he had made it, and

(*iii*) It is conceptually impossible for a person to know his own decision before he has made it.

Since (*iii*) seems to be his most important thesis, I shall here assume, for the sake of argument, that (*ii*) is true (but see below).

---

[10] I regard the dispute over fatalism and free will as a "how-do-you-handle-this-argument?"-game rather than as serious philosophy. (Indirectly it may have a certain philosophical value in pressing forward more precise analyses of "having something in one's power.") So, although it has had a renaissance during the last decade, I shall pay no attention to it in this work. A bibliography including most of the recent articles can be found in Steven Cahn's "Fatalistic Arguments," *The Journal of Philosophy*, vol. 61 (1964), p. 295.

According to Ginet, the possibility of knowing one's own decision in advance involves an absurdity, because the statement "I already know that I shall later decide to do . . ., but I have not yet decided what I shall do . . . that is, I do not yet know what I shall do" (p. 51), makes two logically inconsistent claims.

The main weakness of this argument is Ginet's conception of a decision. He seems to conceive of a decision to do something as some kind of knowledge of something. Once this position is taken, it is rather easy to prove (*iii*), because a person cannot claim both that he has and does not have the knowledge in question. But we must distinguish between deciding and knowing. To decide to kill Olga on May 1 is not to know anything about the future, not even that one will try to kill her. To make a decision is to take a normative standpoint of a certain kind; it is to commit oneself to a certain alternative. Roxbee-Cox, one of the critics of Ginet's position, says aptly "It is in fact rather misleading to emphasize that making up one's mind is a way of gaining knowledge, for this implies that discovering on empirical grounds that a certain decision will be made and actually making that decision are competing methods of gaining the same knowledge" (p. 90).[11]

We may, however, interpret Ginet's article in a different way, viz., as making the point that it is psychologically impossible to deliberate about what to do when one is convinced about what one actually will do.[12] This may be acceptable, though "impossible" is a very strong term. But this psychological difficulty is not a function of the agent's knowledge, but of his beliefs. And even knowledge is all right, if, as pointed out by Roxbee-Cox, the agent forgets what he knows when he deliberates. Moreover, the difficulty has nothing to do with whether our decisions are caused or not.

Suppose Ginet had succeeded in proving that it is "conceptually impossible that a decision should be caused." This would not imply that Smith's decision was not caused, only that *if* it was a decision in Ginet's sense, then it was not caused (Ginet is quite aware of this point). Moreover, he gives no reasons to show that freedom in his sense is relevant to the moral questions of guilt and blame, and this, I think, was a wise omission.

S. Hampshire

Although *Hampshire* (1959), just as Ginet, stresses that a decision involves knowledge of the future, his argument against the possibility of self-prediction does not seem to be based on a cognitive theory of decisions, but is valid, if valid at all, even if a decision is conceived as a normative standpoint of a certain kind.

His main thesis, as far as the present problem is concerned, is that we cannot predict our own decisions without making them. But he admits that "involuntary" decisions, wherein he probably includes decisions based on habits, may be predicted without making them. So what his thesis really amounts to is that we cannot predict those of our decisions which we have it in our power to make or not to make—without actually making them. And perhaps we have no decisions in our power? The argument which we are about to consider does not prove, but presupposes, that we have some decisions in our power.

He formulates his main argument in this way: "Either I take into consideration the reasons that will influence me, in which case I am already engaged in forming a plan; or I somehow contrive to ignore the factors that will influence me, in which case I cannot honestly profess any confidence in my own prediction" (p. 130). His main point, in other words, seems to be that the detailed knowledge of our reasons for and against the different alternatives necessary in order to predict the forthcoming decision presupposes that the reasons are compared and considered in such a way that the decision actually is being made.

Does his argument prove his thesis? I think we can accept that in order to predict my own decision, I must study some "factors that will influence me." But this neither implies that it is necessary to take into consideration my reasons for and against the different alternatives: perhaps it will suffice to study certain neuro-physiological or psychological factors, nor that it is sufficient to do so: perhaps also other factors will influence me.

Knowing about the Prediction

Implicit in the line of reasoning so far considered is the view that the self-predicting agent is not confronted with an outcome which is independent of his own predictive activity in the sense in which the occurrence of an eclipse of the sun is independent of the activity of the predicting astronomer. This view—the view of interaction—is acceptable, and can give the basis for an argument against the possibility of self-prediction which is different from those above. The argument is this: *P*'s prediction at time t_1 of his decision at time t_2 may influence his actual decision at t_2 and thereby the possibility of predicting it. Also another person's prediction may influence him if he gets to know about it, but this other person may have a second prediction in reserve, whereas the chooser necessarily will know his own last prediction.[13]

---

[11] The discussion elicited by Ginet's article includes contributions by, among others, *Canfield* (1961–62), *Swiggart* (1962–63), *Lehrer* (1963), *Thalberg* (1963–64), *Berofsky* (1964), *Taylor* (1964.1).

[12] Some of the psychological problems rooted in the distinction between agent and spectator are discussed by *Hampshire* (1959), *O'Shaughnessy* (1963), and by several authors in *The Monist*, vol. 49 (1965).

A short presentation of this difficulty is given in *Oldenquist* (1963–64). He wants to show that it is impossible to know one's decision in advance. He presents his argument as follows:

(*i*) In order to know at t_1 that I at t_2 will make the decision D, I must, at t_1, know the conditions, say A, B, C, which are causally sufficient for D.

In order to know that A, B, C are causally sufficient for D, I must know that those factors which might lead to *not-D*, will not do so. But my own activity in studying the conditions and predicting my decision may influence the outcome. Hence, I must at t_1, know that it is false that

(*ii*) A, B, C plus my predictive activity, p_1, are sufficient for *not-D*, and, for the same reason, I must know that it is false that

(*iii*) A, B, C plus p_1 plus my knowledge that (*ii*) is false are sufficient for *not-D*.

Hence, we get into an infinite regress. This point, also made by many other authors, seems to me correct, but I do not consider Oldenquist's conclusion well formulated. He says that his argument has shown that "I cannot possibly know that all circumstances besides A, B, C are causally irrelevant to my decision" (p. 56). This formulation does not take us to the heart of the matter, because it refers to a type of uncertainty that we encounter even in experimental science. The crucial thing is that as self-predictors we are always one step behind the mastering of the factors that may be relevant, or as it has been said by Bernard Mayo (in *Mayo* [1958.2]): "no . . . record can . . . include a record of the act of recording" (p. 229).

To assert that our predictions may influence our future decisions does not imply that they always will do so. I see no reason to accept the latter view. The motivations favoring the predicted decision may be so strong that they check whatever counter-influences the prediction might elicit. It is even possible that whereas prediction no. 1 influences the predicted decision so that a prediction no. 2 is needed, no. 2 does not modify the predicted result. Whether a decision will be influenced by one's own final prediction is an empirical question,[14] but if such influence takes place, the prediction will be falsified.

Conclusion

The authors presented have focused their attention on the possibility of self-prediction. With this as a starting-point, Ginet tried to prove that our concept of a decision implies that a decision is uncaused, Hampshire that we cannot know our decisions in advance, and Oldenquist that predictive knowledge is impossible because of the effect of knowing about the prediction. I have been more or less critical toward their whole approach. It may be of value from the point of view of obtaining a better understanding of some of our concepts or of the agent and his situation, but it is wasted from the point of view of solving the question of free-will. This is easily seen if we assume that they had succeeded in proving that self-predictions are impossible, and that this discovery was communicated to Mr. Smith's judge, who then replied to Smith's excuse by saying "I am sorry Smith, but you are to hang, for it has been discovered that your will was free in the sense that it was impossible for you to predict in advance that you would decide to kill Olga." A reply which is either nonsensical or cruel or both.

As mentioned above, few authors deal directly with the question whether or not our decisions and actions are caused.[15] Nevertheless, some words about this issue seem appropriate.

---

[13] As regards the implications of knowing about the prediction, see also *Lyon* (1959), *MacKay* (1960), *C. J. F. Williams* (1960–61), *Canfield* (1961) & (1961–62), *Beck* (1965), and *Perry* (1965).

[14] Cp. *Stenner* (1964).

[15] But see, e.g., *Farrer* (1958) and *Ofstad* (1961). A discussion of different interpretations of "determinism" and different arguments pertaining to this issue will be found in *Hook* (1958), Pts. I & II. See also *Brandt* (1959), *Hospers* (1961), and *Pears* (1963), ch. 4. The use of causal explanations in science is discussed in *Bohm* (1957), *Grünbaum* (1957), *Scriven* (1957), *Bunge* (1959), *Waismann* (1959), *Munn* (1960), *Nagel* (1961), and *Boring* (1963).

So far neither psychological nor physiological research have succeeded in establishing any universal law of relevance to the present discussion. Moreover, it seems very likely that at least some characteristics of some decisions and actions will remain statistically rather than universally related to other factors in the universe. One reason for this: the interaction between the predictor and the phenomena to be predicted has already been mentioned. Another reason is the limited intersubjective accessibility of the required data. In some cases it may be the morally most important characteristics which are only statistically related to other factors, in other cases these characteristics may perhaps be universally related to other factors.

Let us suppose that we knew that those characteristics of Smith's decision which made it morally wrong had a low degree of statistical relation to other factors in the universe. Would this give us a basis for an answer to his excuse? We could say "Smith, you are to hang, for your will was free in the sense that those characteristics of your decision which made it morally wrong had a low degree of statistical relation to other factors in the universe." But the answer is ridiculous. And it is ridiculous because it is just as irrelevant as if we said "Smith, you are to hang, because your hair is green."

V. Two-Domainism

According to Kant, our decisions and actions are causally determined *qua* appearances, but neither caused nor uncaused *qua* things in themselves. The logic of this attempt of solving the alleged dilemma between determinism and free will: splitting the world into two domains, one subjected to causality, another beyond it, has also some modern followers,[16] of which *Melden* (1961) perhaps is the most emphatic one.[17]

Applied to our example, Melden's view may be expressed thus: Smith's bodily movements, e.g., his pressing the trigger, can be described in behavioral or neuro-physiological terms and are causally determined (he takes the term "cause" in what he calls a "Humean sense"), but Smith's action—his killing his aunt—although involving bodily movements, is not identical with a bodily movement, and neither caused nor uncaused.

I shall limit my discussion to three questions: (*i*) What is, according to Melden, an action "more than" a bodily movement? (*ii*) Why is the causal model inapplicable to actions? (*iii*) Does his line of thinking have any significance for the question of responsibility and freedom?

Melden tries to answer (*i*) by analyzing the example of a driver signalling at a cross-road. The driver moves his arm in a certain way, but "because of our familiarity with the rules of the road, we recognize . . . that he is signalling" (p. 191). The report that the driver is signalling is not based on different sense-impressions from those we have when observing his bodily movements. No telescope is needed to see that the driver is signalling and not only moving his arm. Melden says, quite explicitly, that although "moving the arm" and "signalling" are different descriptions, a case of the former is, in appropriate circumstances, the "very same thing" as a case of the latter. So what is the difference? His point seems to be that the observer seeing the movement of the arm as a case of signalling has been trained through certain social rules. He meets the stimuli with an acquired ability to understand what he observes, just as he meets certain geometrical figures on paper with an ability to perceive them as letters. Melden's idea, in other words, seems to be that P's bodily movement M in the situation S in society X at time t, is an instance of the action A if and only if there is within X, at t, a rule according to which movements such as M, should, in S, be recognized as actions of type A.

But how shall we understand the term "action" here? What is it that the social rules help us to recognize the bodily movement as? As an action? That is, as a bodily movement recognized as . . .? The concept of an action remains unexplained. Moreover, social rules are not even infallible instruments for the identification of actions, for whereas the driver according to the rules of the road was signalling, actually he was perhaps pointing to a church, or his arm was made to move by an electric impulse transmitted to his brain by an evil spirit. Furthermore, a person may act even if he is a nonconventional actor doing something which we have not been trained to understand. He may have to explain the nature of his action to the puzzled observers. I find Melden's view unacceptable, and first of all because he puts the rules of society where he should have put the intention of the individual.[18]

---

[16] See in this connection also *MacKay* (1960) and *Körner* (1964), but their views, different from each other, are not based on the kind of arguments which here will be considered.

[17] Critical comments on Melden will be found, e.g., in *Ewing* (1963–64) and *Hamlyn* (1964). Cp. also G. J. Warnock's "Actions and Events" in *Pears* (1963).

Melden tries to establish (*ii*), not by proving it, but by trying to show that our usual explanations of actions are of non-causal nature. For instance, to explain Smith's killing by saying that he had decided to kill, is not a causal statement, for his decision is intelligible only as a decision to kill.

Very seldom, however, do we explain actions with reference to decisions, but forgetting about that, the important thing is that his argument is fallacious. Even if Smith's decision is described as a decision to kill Olga, his decision may or may not be followed by action. Perhaps he died just after his decision, or someone held his hand, and even if he acted, he may not do what he decided to do. And even if he did what he decided to do, the empirical relation between his decision and action may exemplify a universal relation between decisions of a certain type and actions of a certain type. Finally, even if Melden had been right on this point, it would have revealed only some features of our common ways of explaining actions. It would not have shown that a causal model is inapplicable to actions.

Let us assume, for the sake of argument, that Melden had succeeded in proving that actions are beyond causality. Would this have any significance for the free-will issue? Melden does not deny that Smith's bodily movements are causally determined. Hence, if the judge says (basing himself on Melden) "I hold you morally responsible because your action was neither caused nor uncaused," Smith can answer (basing himself on Melden): "Yes, that is true, but my killing involved a set of bodily movements and they were causally determined. They could not have been different from what they were. Hence, your condemnation is unjustified."

Melden, I am happy to say, has not tried to show that the statement that an action is neither caused nor uncaused implies that the agent had it in his power to act otherwise. So even if the thesis of his book were true, he would not have shown that we are free in the sense required by moral guilt or blame.

VI. Reason Versus Cause

Hampshire's and Melden's studies of intention, choice, action, voluntary action, reason for an action, etc., are parts of a general revival of so-called philosophical psychology. Let us look at two of the points which often are made by recent writers within this tradition, and let us formulate them so vaguely that we can use them as starting-points for different interpretations: (*a*) a decision or action is not a definite entity, and (*b*) we decide and act on the basis of reasons. Let us limit our discussion to two questions: (*i*) Does (*a*) or (*b*) have any consequences for the question whether or not our decisions or actions are caused? (*ii*) Does (*a*) or (*b*) have any consequences for the question whether or not we have it in our power to decide or act otherwise?

I start with (*a*) and distinguish three interpretations, formulating them in terms of our example.

(a_1): even the most minute observations of Smith's behavior is compatible with different hypotheses about what he was doing. His action may have been killing Olga, but it may also have been trying-to-find-out-whether-the-gun-is-loaded-or-not, or trying-to-find-out-whether-or-not-the-bullet-would-go-through-her-heart, or getting-rid-of-an-obstacle-in-order-to-get-hold-of-Olga's-cat.

(a_1) seems acceptable, although observations of a larger segment of Smith's behavior may certainly limit the class of plausible classifications. But since I cannot see that (a_1) has any consequences of the type indicated in (*i*) and (*ii*), I shall say no more about it.

(a_2): the statement that Smith killed Olga is not an empirical statement.

(a_2) is false. Whether Smith killed Olga or not depends, among other things, upon his behavior and intentions, and to decide what he intended is an empirical task.[19] But even if (a_2) were true, it would be irrelevant to the question of Smith's power to have acted otherwise. On the other hand, if we take the term "caused" in such a way that only empirical events can be caused, and interpret (a_2) to imply that an action is not an empirical event, then it follows that actions are not caused.

(a_3): Smith's action can be described in different ways. We can say "Smith killed Olga" or "Smith shot Olga" or "Smith shot Olga with his red gun," etc.

(a_3) is acceptable, but has no implications for the question whether or not his action was caused. But could we say that the possibility of alternative descriptions gives a wonderful solution of the question whether or not Smith had it in his power to act otherwise? He did shoot her, but he could have acted otherwise, for he did have it in his power to do the action describable by the sentence "Smith shot Olga with his red gun." But this must be a joke.

---

[18] Cp. *Yolton* (1962).

[19] Cp. *Geach* (1960).

Let us turn to (*b*). Does the statement "Smith had a reason for killing Olga" imply or exclude that "Smith's killing Olga was caused"? That depends, first of all, upon what we mean by "reason." If by Smith's "reason" we refer, *inter alia*, to his pro-attitude toward killing Olga (cp. Davidson, p. 687), it is rather obvious that to explain Smith's action in terms of his reasons may be an example of a causal explanation. On the other hand, if by Smith's "reason" we refer to Olga's death, saying that the fact that she would die if he pulled the trigger, was the reason why he pulled the trigger (cp. Shwayder, p. 86), then a reason is not a cause, although Smith's anticipation of the fact in question may have been a causal factor. Similarly, if by Smith's "reason" we refer to what he thinks justifies his action ("She always was so angry with her cat"), his reason is not a cause, although his desire to improve the cat's conditions may have been a causal factor. Finally, if "reason" refers to the goal toward which his action was a means (e.g., to become a cat-owner), then, again, a reason is not necessarily a causal factor, but Smith's belief that shooting would help him to realize his goal might be. Hence, we sometimes use "reason" in such a way that a reason is a causal factor, but even when we do not use it that way, the statement that an action was done for a reason does not imply that it was not caused.[20]

Does the fact that we act on the basis of reasons have any consequences for the question whether we have it in our power to act otherwise than we in fact do? A plausible answer seems to be that we do have a certain degree of power both over the amount of positive or negative value that we attribute to a certain anticipated consequence as well as over the weight we give this value as compared with other values. By modifying our evaluations, or by trying to see our decision or action within a wider perspective of norms and values, we may increase our power to decide and act in different ways.[21]

VII. Moral and Neutral Libertarianism

C. A. Campbell

Campbell is one of the most energetic defenders of the so-called libertarian position both before and within the period selected for this article. His position can be summarized as follows:

(*i*) The crucial question is not whether there is some sense in which we are free, but whether we are free in the sense required in order that we shall "deem it proper to attribute moral praise or blame to agents" (1957, p. 159).

(*ii*) In order to find this sense, we must analyze what Campbell calls the "moral consciousness."

(*iii*) Such analysis shows (*a*) that the crucial question is whether our choices and decisions (our "inner acts") are free or not, because even if external circumstances prevent a morally wrong act, we may be responsible for having decided to do this act, and, on the other hand, we are not responsible for an action unless it is rooted in an "inner act," for we do not consider robots responsible. It also shows (*b*) that in order that we shall be proper objects of moral praise or blame, our decisions must be free in the sense that the two following conditions are fulfilled:

(1) the sole cause (author) of the decision is the agent himself, and not his character as formed by heredity and environment (p. 160).[22]

(2) the agent could have decided otherwise than he in fact did, in a contra-causal sense of "could" (p. 164).

Campbell does not maintain that every decision is free in this sense. Our decisions to follow this or that desire are caused and hence not free in the sense indicated. It is only the choice between following one's strongest desire and doing what one believes one ought to do, which is causally undetermined and has the self as sole author. In his own words: "In the act of deciding whether to put forth or withhold the moral effort required to resist temptation and rise to duty,[23] is to be found an act which is free in the sense required for moral responsibility" (p. 168).

Campbell's main reason for maintaining that we are free in the sense indicated is that an analysis of our "moral consciousness" shows that it is impossible for us to disbelieve that we are free in that sense. I find it hard to take this line of argumentation seriously, but since its weaknesses have been pointed out by many authors,[24] I shall say no more about it, but rather consider two other questions: Does freedom in the sense indicated suffice for moral responsibility? Has Campbell formulated the sense of "free" which he himself seems to consider crucial?

---

[20] Cp. *Ayer* (1963), pp. 266 ff.

[21] Cp. *Hampshire* (1959), p. 129 and *Ofstad* (1961), pp. 196–203.

[22] Also Chisholm stresses, in his lecture (1964), the role of the self in human freedom.

[23] Campbell says that this means "doing what one believes we ought to do." See (1957), p. 173.

First of all, his position seems to imply that a wrongdoer who did not think about how he ought to act cannot be morally responsible (p. 221)—a very strange consequence. Secondly, Campbell seems to postulate a very odd principle of causality, for it withdraws in the moment the agent thinks about his duty and reappears as soon as he forgets about it. Thirdly, even a person who thought about what he ought to do is not responsible if he did not decide to put forth or withhold effort. For if he did not, then he did not do the act which Campbell considers free. Fourthly, let us suppose that Eichmann deliberated about whether to have the Jews shot, A_1, (which, we assume, was the solution he desired) or killed by gas, A_2, (which, we assume, he considered his duty), and that he decided to do what he believed to be his duty. Let us also assume that his decision was free in Campbell's sense. Does this imply that it was free in the sense required in order that he should be morally responsible? It does not, because it does not imply that Eichmann had it in his power to do something which would have been less evil than A_1 or A_2. Fifthly, let us suppose that if Eichmann had decided to save the Jews, they would have been saved. Even this does not show that he was morally responsible, for perhaps he did not have it in his power to believe that it could be right to act contrary to Hitler's order, and even if he did have this in his power, perhaps he did not have it in his power to act on the basis of this belief. Finally, even if "Eichmann himself" was the "sole cause" of his decision, how can we avoid the consequence that also that part of Eichmann which was formed by heredity and environment will experience our reactions?[25]

Campbell's philosophy of freedom seems to have been formulated by a person living within an unproblematic moral universe where the paradigm-case of potential wrong-doing is: Shall I steal a little bit of mother's jam or shall I "rise to duty"?

Let us turn to the second question. I do not think Campbell really means that a decision to put forth or withhold effort is free only in the sense he indicates, for this sense does not exclude that it is a matter of luck whether the effort will turn out to be of sufficient strength. He probably also means that the agent is free to mobilize the amount of effort required in order to resist any temptations (p. 221). But if this is what he means, then he should have argued that we have it in our power to put forth this amount of effort, and not that we have contra-causal freedom, for freedom in the latter sense (no matter its exact meaning) does not exclude that the agent was powerless. Moreover, it is hard to reconcile his contra-causal sense of "could" with his belief that it may be more or less difficult to "rise to duty" (p. 220). Also this belief suggests that he has in mind a comparative concept of power to decide. My guess, therefore, is that a definition of "freedom" in terms of power to decide would have been a more adequate expression of his intentions.

R. Taylor

Contrary to Campbell, Taylor in his contribution to *Hook* (1958), and in his book (1963), does not limit our freedom to situations where we choose between following our strongest desire or doing what we think we ought to do. But he shares Campbell's view that causal indeterminacy is necessary but not sufficient for freedom in the sense required by moral responsibility. It is also necessary that the act was an act of our self. By "self" Taylor seems to refer to a transempirical entity, and the statement that we have such a self cannot be subjected to empirical tests. But why should a person be morally responsible if his act was rooted in this transempirical entity, and without responsibility if it was caused by his very substantial desire to kill?

Let us assume, for the sake of argument, that Smith is a self of the kind Taylor seems to have in mind, and that this self "originated" the act of killing (A_1). This does not imply that Smith had it in his power not to kill Olga. Perhaps Smith's self originated A_1 with such a strength that he was unable to act otherwise. Moreover, if Smith was powerless in so far as A_1 was caused by Smith's "self," defined, *inter alia*, in terms of his thoughts and desires, how can we exclude the possibility that he is just as powerless if we define him in terms of a certain transempirical or substantial self? Must we not postulate a super-substantial self to prevent that Smith becomes a slave of his substantial self? And a super-super-substantial self to . . . , and so on? And if Taylor answers that the objection is illegitimate, because Smith = his substantial self, why not avoid the whole mystery by setting Smith = his empirical self?

---

[24] Campbell's way of thinking was defended by *Lehrer* (1960). Criticism can be found in *Bradley* (1958), *Franklin* (1962), *Nowell-Smith* (1954). Campbell's answers will be found in (1958), (1962) and to *Nowell-Smith* in (1957), pp. 214–228.

[25] Obviously, my use of this example does not pretend to correspond to the historical facts.

Indeterminism, Libertarianism, and the Belief in Transempirical Entities

The best working hypothesis when dealing with an author maintaining that moral responsibility presupposes freedom in the sense that the agent's decision was uncaused or unpredictable is that his thinking is confused. Probably he has failed to see the difference between the following two interpretations of "P, in S, could have decided D_2 instead of D_1": (*i*) It was equally probable that P, in S, should make the decision D_2 as D_1, and (*ii*) P, in S, had it in his power to decide D_2 instead of D_1. The latter statement tells us something about the agent. The former tells us something about D_1 and D_2's relations to other events in the universe. Failing to distinguish between them, the author may try to prove (*i*), intending to prove (*ii*). But there is also another and deeper reason why a philosopher may defend indeterminism: he may be dissatisfied with our empirical self or our limited power and postulate the existence of a transempirical self or power-center, defending indeterminism in order to exclude the applicability of Occam's razor.

Campbell does not use the expression "transempirical power-center," but his unwillingness to accept empirical disconfirmation of the statement "P had it in his power to decide D_2 instead of D_1" suggests that he postulates something like such a power-center. Taylor does not use the expression "transempirical self," but it is rather clear that he refers to a transempirical entity. He seems unwilling to evaluate the empirical self highly enough to identify it with one's real self. Given this situation, libertarians, such as Campbell and Taylor, may come to defend indeterminism to meet the following objection of the scientist: "I don't know whether we have a transempirical self or power-center, or not. I don't even understand what these expressions mean. I know only that I don't have to worry about such things. I can predict the characteristics of our decisions without paying any attention to the operation of any such entities. And I must be permitted to say, therefore, that this self or power-center seems to live a rather quiet life." My hypothesis is that the libertarian will try to meet this objection by denying that the scientist can give perfect predictions. Such predictions are impossible, he may maintain, because the transempirical entity manifests itself in the empirical domain in a way which for the scientist must appear as partly random. In other words, my hypothesis is that the libertarian's indeterminism is intended as a wall of defense around his belief in certain transempirical entities.[26]

VIII. Hard Determinism and Hard Evolutionism

The position of the hard determinist consists of three standpoints: (*i*) he asserts that our actions and decisions are parts of causal sequences (ancestorially caused), (*ii*) he has a certain view on the morally correct or "proper" sense of "moral responsibility," and (*iii*) he maintains that moral responsibility in this sense requires freedom in a sense which is incompatible with (*i*).

I have earlier distinguished two variations of hard determinism. According to the first, the principle of causality is incompatible with the moral aspect of guilt and blame, but we may be justified in holding people guilty and blaming them in a non-moral sense. According to the second, the principle of causality is incompatible with the notions of desert and merit, but we may be justified in making use of moral as well as non-moral blame on utilitarian grounds.

Recently hard determinism of the first kind has been defended, among others, by Paul Edwards in his contribution to *Hook* (1958).[27] He does not deny that the principle of causality is compatible with freedom of action, if by calling an action "free" we mean only that it may be influenced through the use of blame (limiting ourselves to wrongdoings). He does not even deny that we use the expression "moral responsibility" in such a way that it requires freedom in this sense only. But he maintains that this is an improper sense, and that if we take "moral responsibility" in its "proper" sense, then the principle of causality implies that we are not morally responsible. The crucial questions, then, are what Edwards means by "moral responsibility in the proper sense," why he holds this to be the proper sense, and why the principle of causality excludes responsibility in this sense.

---

[26] Cp. at this point *Ofstad* (1955), (1961), pp. 232–234, and (1963).

[27] *Baylis* (1958), pp. 37–38, seems to take the second position, but I shall not go into his reasoning here.

Edwards gives no explicit answers to the first and second question, but he seems to mean that a judgment about guilt or blame, to be moral in the proper sense, must be independent of violent emotions and based on general principles.[28] Let us accept this and let us turn to the problem which is most relevant from the point of view of this article: Why does the principle of causality exclude that we can be morally responsible in the "proper" sense.

Edwards' answer, applied to our example, can perhaps be thus expressed: In order that we shall be justified in morally blaming Smith for killing Olga, he must have had it in his power to act otherwise. To have had such power, he must have chosen his own character (p. 123). But the principle of causality implies that our character is formed by heredity and environment. And since this principle must be accepted, it follows that we are not justified in morally blaming Smith.

This reasoning seems to me confused.[29] First of all, the principle of causality implies at most that Smith's character is caused, not that it is caused by heredity and environment. Secondly, even if it did imply this, it does not follow that Smith was unable to abstain from killing. Thirdly, even if it followed that Smith was powerless in this respect, it does not follow that he would be less powerless if he had chosen his own character. For the fact that Smith in the year 1900 chose his character does not guarantee that he on May 1, 1965, had it in his power to abstain from killing. Moreover, what does it mean to "choose one's own character"?[30] And who chose the character of that Mr. Smith who in 1900 chose the character of Mr. Smith?

Let us make the plausible assumption that Edwards requires that Smith in 1900 not only chose his character, but that he also had it in his power to choose another character than the one he actually chose. Then this gives us a sense of "having something in one's power," and if we apply this sense to the situation where Smith killed Olga, it is conceivable that in this sense he had it in his power not to have killed her; and we have thereby reached the crucial question: Whether Smith on May 1, 1965, had it in his power to act otherwise. No power over his character or over anything else can help if he lacked the required amount of power in that situation.

Hence, the critical thing for the hard determinist is to prove that there is a moral-responsibility-relevant sense of "having something in one's power" which is such that the agent's power to have abstained from that for which he is held responsible (a decision, an action, etc.) is reduced or annihilated if we assume that his decision or action was part of a causal sequence. Edwards' attempt to show that we have not chosen our own character—even forgetting about what this really can mean—misses the point.

Whereas Edwards tried to reach his conclusions by arguing from so-called deterministic premisses, other philosophers, like Hospers, have tried to reach similar conclusions from premisses referring to hereditary and psychological factors (cp. *Hospers* (1961) and his contribution to *Hook* (1958)).[31]

Hospers distinguishes between what he calls an "upper" and a "deeper" level of moral discourse. If we conduct our discussion on the former level, we are justified in considering people free in so far as they would have acted otherwise if they had wanted to, and to make use of moral blame in so far as this helps to make them do what they ought to do. But to talk about deserving moral blame, we must move on to a "deeper" level of discourse. Hospers' position seems to be that we would deserve blame for our wrongdoings only if we had caused our own character. But neither Smith nor anyone else is free in this sense, for our characters are formed by hereditary and environmental factors.

The arguments which showed that Edwards' view is untenable show also that Hospers' view is untenable. Rather than repeating my arguments, I shall comment on an important difference between their views. Whereas Edwards means that we are never free and responsible in what he calls the "proper sense," Hospers, at some places in his article, seems to mean that we may be free in such a sense that we deserve moral blame.

---

[28] He seems to mean that theories stressing modifiability must be based on emotions—not reasons—but he gives no reasons for this.

[29] *Pitcher* (1961) criticizes hard determinism, arguing that it does not distinguish between the theses that some causes excuse and that all causes excuse.

[30] Hook, in his contribution "Necessity, Indeterminism, and Sentimentalism" to *Hook* (1958), points to this unclarity, see pp. 187–188. And Pap maintains that it is just as meaningless to talk about "choosing my own character" as "eating my own character." See "Determinism, Freedom, Moral Responsibility, and Causal Talk" in *Hook* (1958), p. 215.

[31] A somewhat similar point of view has been developed by *Matson* (1956).

Let us assume that Smith on May 1, 1965, had it in his power to overcome his urge to kill. According to Hospers, he does not deserve praise for this, for it is, just as his character, a product of hereditary and environmental factors. But once he has this power, what shall we say if he does not make use of it? In that case, Hospers says, moral exhortations may be useful. But I think he means more than that, for the utility-point-of-view belongs to the "upper level," whereas Hospers, when he makes this point, is discussing at the "deeper level." He probably means that in this case Smith deserves moral blame. Does this imply that he means that Smith in this case has caused his own character? I think it rather indicates that he has moved on to a different, and, morally speaking, more relevant meaning of "free," viz., whether or not one was able to—had it in one's power to—overcome one's urge to kill. If we interpret his view in this way, his stress on the traumatic effect of early childhood experiences may be seen, not as an alleged proof that we are never free in the sense required by moral responsibility in the so-called deeper sense, but as emphasizing some of the factors which may lead to a reduction of our power to decide and act as we ought to.

Some hard determinists consider their premisses as true *a priori*, and maintain, therefore, that the conclusion that we are not free in the sense required by moral responsibility, is also true *a priori*. The hard evolutionist, on the other hand, argues on the basis of empirical statements and can prove at most that it is very unlikely that we are free in the required sense.[32] However, if he takes as his premiss the empirical statement that we have not caused or chosen our character, or that it is caused by early childhood experiences, then he can prove his thesis only if he can establish an empirical relation between, e.g., some childhood experiences and, on the other hand, the adult's lack of power to decide otherwise. Empirical research of relevance to the establishment of such relations has barely started.

IX. MILD DETERMINISM

If "mild determinism" refers to the view that the principle of causality is compatible with freedom in the sense required by moral responsibility, then an indeterminist can be a mild determinist, and this is a confusing terminology. Hence, I shall require that a philosopher, in order to be called a "mild determinist," accepts that our decisions and actions are parts of causal sequences.

In what sense of "morally responsible" does mild determinism maintain that we can be morally responsible? Some of them seem to take this expression as meaning that behavior is modifiable through the use of praise and blame. But according to this usage, hard determinism implies mild determinism. That mild determinism shall be different from hard determinism in an interesting way, the mild determinist must maintain that the principle of causality is compatible even with freedom in the sense required in order to be morally responsible in the sense accepted as the "proper" sense by the hard determinist, e.g., in the sense that one can be morally guilty or blameworthy. But not all mild determinists make this contention. One reason for this may be that they consider such notions as moral guilt or blameworthiness as logically contradictory. Another reason may be that the mild determinists agree with the hard determinists that causality is incompatible with freedom in the sense required by moral guilt or blameworthiness, but conclude, in opposition to the hard determinists, that this shows that these notions are unacceptable, and that the correct notion is modifiability. The difference between them would then be that the mild determinist asserts, and the hard determinist denies, that modifiability can be a morally relevant sense of "responsible."[33]

In the following I shall limit my discussion to the reasoning of those who at least appear to maintain that the principle of causality is compatible with freedom in the sense required in order to be morally guilty or blameworthy.

P. H. Nowell-Smith

Nowell-Smith, one of the ablest defenders of mild determinism of the last decade,[34] does not define a notion of moral guilt, but looking at the list of sentences whose presuppositions he intends to discuss (1954, p. 270), we seem justified in interpreting his view in the following way:

---

[32] Hook, it seems to me, is quite right in criticizing some hard determinists for oscillating between holding that it is self-contradictory to assert that we are free in the sense required by moral responsibility, and that we seldom, if ever, are free in the required sense. *Op. cit.*, pp. 187–188.

[33] Cp. *Smart* (1961).

[34] Arguments in favor of this view can be found also in, e.g., Blanshard, *op. cit.*, Ducasse "Determinism, Freedom and Responsibility" in *Hook* (1958), Hook, *op. cit.*, Pap, *op. cit.*, *Brandt* (1959), *Mandelbaum* (1959–60), *Ofstad* (1961), ch. VI, sect. 18, *Zink* (1962), and *Vivian* (1964). Cp. also *Hampshire* (1965).

(*i*) Smith may be morally responsible for killing Olga both in the sense that he was morally bad in doing what he did (cp. my sense of "moral guilt"), and in the sense that he was blameworthy.

That someone shall be justified in holding Smith responsible in these two senses, two conditions must be fulfilled:

(*ii*) In killing Olga, Smith did something morally wrong.

(*iii*) Smith does not have an acceptable excuse for what he did.

Smith may excuse himself in different ways, but I shall limit myself to Nowell-Smith's discussion of the excuse that Smith did not act freely. The morally relevant sense of "free" is, according to Nowell-Smith, that the agent could have acted otherwise. If Smith could not have acted otherwise, then he is not guilty and not blameworthy. If he could have acted otherwise, then he is guilty and blameworthy, unless he has another acceptable excuse. The questions are then: How does Nowell-Smith interpret "Smith could have acted otherwise" (T)? Is it true that freedom in this sense is compatible with the principle of causality? Does freedom in this sense suffice for guaranteeing that Smith was free in the sense required by guilt and blameworthiness?

According to Nowell-Smith,

(*iv*) T is equivalent to "Smith would have acted otherwise, if this or that condition had been fulfilled" (T_1).

But it is always possible to find a substitution which will make T_1 true. Hence, we need some criterion to distinguish between permitted and forbidden substitutions, otherwise Smith can never excuse himself by saying that he was not free. Before looking at Nowell-Smith's criterion, let us, however, present his view as regards what kind of statement T is and how it can be verified and falsified.

His view seems to be that T cannot be directly verified or falsified by observing what Smith did, for the fact that he killed Olga neither proves nor disproves that he could have acted otherwise. To establish T we must show that Smith had the capacity to abstain from killing and that nothing in the situation prevented him from making use of his capacity. To disprove T it suffices to show that he either lacked the capacity or that something prevented him from making use of it. How can we prove that Smith had the capacity and that nothing prevented him? Nowell-Smith's view seems to be that positive as well as negative evidence for this statement must be based upon observations of Smith's behavior in other situations. For instance, we must try to establish that Smith in earlier situations has been able to overcome impulses to kill. Positive evidence of this kind is always open to the objection: "How can we be certain that this shows that he "could have abstained on May 1, 1965?" and negative evidence is open to a similar objection: "How can we exclude the possibility that he might have been able to abstain on May 1?", but, according to the author, it would be absurd to base responsibility, as well as the absence of responsibility, on the logical possibility of such contingencies.

Let us now return to the question of how Nowell-Smith distinguishes between those if-substitutions which exclude guilt and blameworthiness and those which do not. His view on this point may be interpreted in different ways. The most plausible interpretation seems to me to be that

(*v*) If T_1 is true only for such if-substitutions which the moral rules accept as excuses, then Smith can excuse himself.

For instance, moral rules accept unconsciousness as an excuse, and if Smith was unconscious when he killed Olga, so that T_1 is true only if he had not been unconscious, then he has an acceptable excuse. But if he would have acted otherwise if he had tried to, then he is responsible, for the moral rules do not accept not-trying as an excuse. Nowell-Smith mentions "acted under physical compulsion," "acted under ignorance of fact" as examples of excuses which the moral rules accept, and "not wanting to do what is right," "being too dishonest," as typical examples of causes which the rules do not accept as excuses.

He tries to explain why moral rules accept only certain special conditions as excuses. His explanations seems to be that

(*vi*) Moral rules accept as excuses only conditions of such a kind that if wrongdoers acted under such conditions, then, as a general rule, it will be pointless to blame them.

If Smith would have acted otherwise only if he had not been unconscious or subjected to physical compulsion, then he has an acceptable excuse because it is, as a general rule, pointless to blame persons who acted wrongly for such reasons. And it is pointless because the purpose of blaming is to strengthen or weaken certain traits of character.

Simplifying his view as much as I can, it seems to me to boil down to this:

(*vii*) Smith's act in S is morally wrong and free if and only if Smith's act is in conflict with a moral rule, and it is, in general, possible to discourage people from acting in the way Smith did by blaming them. (Let us express this by saying that the act is *modifiable in general*).

If (*vii*) is fulfilled, then Smith acted freely and is guilty and blameworthy. And he is guilty and blameworthy even if blaming him will increase the likelihood that he will kill again. On the other hand, if Smith's act is not modifiable in general, then he was not free and is neither guilty nor blameworthy, even if blaming Smith would have had good effects on his character.

I find these consequences unacceptable. The emphasis on general modifiability may be convenient within legal systems, but in moral contexts we tend to hold that a person is not guilty unless he—the specific person in question—had it in his power to act otherwise. Nowell-Smith thinks about moral problems in a too legalistic way. He talks about moral rules as if they constituted a definite group of rules roughly as legal rules. And he does not clearly distinguish between the conditions which certain rules recognize as valid excuses, and the conditions which ought to be accepted as valid excuses.

Nowell-Smith would probably object to this criticism by saying that his requirement that the agent could have acted otherwise refers to what the person in question was able to do in a particular situation. But this line of defense is excluded, because it is the analysis of this requirement which has led to (*vii*).

At this point it may be clarifying to distinguish between a general and a particular use of "could" and "power"-sentences.[35] If we take "could" or "power" in a particular sense, saying that a person could or had it in his power to do A_2 instead of A_1, in S, means that he in that particular situation was fully able to do A_2. The fact that he never before has been able to do so is not logically incompatible with the assertion. The fact that he always has been able to do so does not logically imply it. We are considering one isolated moment. If we take "could" or "power" in a general sense, saying that a person could or had it in his power to do A_2 instead of A_1, in S, means that he generally is able to do such actions as A_2. In this sense the agent could have acted otherwise in S even if he was unable to act otherwise in the particular sense. Using this terminology, my view can be expressed by saying that it is the particular and not the general use of "power" which is relevant to questions of moral guilt and blame. The general use of "power" is of significance only indirectly, as evidence for or against the possession of power in the particular sense.

In his answer (1960) to J. L. Austin's criticism of his view (1956), Nowell-Smith seems more aware of the distinction between the particular and the general use of "could" than in *Ethics*, and he raises the question whether we can exclude the possibility that a person who usually is able to do something may lack the ability "just then and there" (p. 97), but he does not discuss the modifications which would have to be made in his earlier view if we accept the particular use of "could" as the morally relevant one.

This brings me to another point. Both in *Ethics* and in his answer to Austin, Nowell-Smith is too much influenced by the idea that the wrongdoer must have had an ability to act otherwise. But to say that he must have been able to act otherwise is not the same as to say that he must have had an ability to do so.[36] Smith may have had all the abilities pertaining to A_1 and A_2 that a person can have, but nevertheless have been powerless, because it was a question of resisting a temptation, and to resist a temptation is not an ability. It may be granted that Smith's situation may be similar to the situation of the frustrated golfer in the sense that both tried to do something and failed. But their situations are very different. It makes good sense to say "I am taking a course in how to hole six-foot putts. I am doing quite well. Very soon I shall usually succeed when I try, and then I shall obtain my diploma for having acquired an ability," but the following would be quite strange: "I am taking a course in how to resist temptations to kill aunts. I am doing quite well. Last week I resisted six out of ten temptations. As soon as I can resist eight out of ten (if there are any aunts left at that time!), I shall get my diploma for having acquired an ability." It is misleading to say that moral guilt or blame presupposes the ability to act otherwise. This is one of the reasons why I prefer to say that guilt presupposes that the agent had it in his power to act otherwise.

As indicated above, Nowell-Smith points out that T can be proved or disproved only through inductive evidence, which, of course, is true. But when he classifies persons emphasizing the uncertainty of such evidence as ordinary sceptics of induction, he misses an important point, viz., that the uncertainty of the evidence primarily is due to the difficulty of getting hold of relevant evidence. That Smith's record of previous not-killing shall count as evidence for the hypothesis that he had it in his power to resist his urge to kill Olga, we must know that the previous cases are similar in all relevant respects to the present one. And we do not know this, and this lack of knowledge is not a consequence of the inductive character of the evidence.

---

[35] Cp. *Honoré* (1964).

[36] Cp., e.g., *Locke* (1962) and *Thalberg* (1961–62).

The answer to the second question I raised in the beginning of my discussion of Nowell-Smith's view, whether or not freedom in his sense is compatible with the principle of causality, obviously is in the affirmative. Let us turn to the third, and more interesting question: whether freedom in Nowell-Smith's sense gives the agent the freedom required by moral guilt and blame.

He does not give a clear answer to this question. He begins his discussion by distinguishing between guilt, blameworthiness, and blame, but he does not discuss the presuppositions of guilt and blame separately, and in the end the concept of individual moral guilt disappears completely. His view seems to be that a person is guilty if and only if he has disobeyed a moral rule and none of the excuses admitted by the rule applies to him, and this means that a person is guilty if and only if he has acted contrary to the moral rules. And this is legalistic thinking about moral problems.

What about free will as a condition of blame? In several of his publications Nowell-Smith has charged philosophers for introducing concepts of freedom whose moral relevance is not clear. As opposed to them, he maintains that his own sense: "*P* would have acted otherwise, if this or that condition had been fulfilled" can be justified as morally relevant, because it is pointless to blame unless the agent was free in some such sense. But, as shown above, his requirement boils down to a requirement about general modifiability. And blaming Smith may discourage him from future wrongdoings even if the act in question is not modifiable in general, and, even if it is, blame may have no effects on Smith. Suppose we then interpret his requirement in such a way that it suffices that Smith's future behavior will be modified. Saying that freedom in this sense can distinguish between exculpatory and non-exculpatory cases, is, however, not very interesting, since it becomes analytically true. And if we interpret his requirement to be that the agent would have acted otherwise, if he had tried to act otherwise, his justification disappears, for it may be useful, both in general and in the particular case, to blame a person even if he would not have acted otherwise if he had tried to, and it may be useless even if he would have acted otherwise if he had tried to.[37] The requirement that Smith had it in his power to act otherwise, may, however, come in as an empirical requirement: the likelihood of producing good effects through the use of blame may be higher if we blame only persons who in some sense had it in their power to act otherwise. But once we have raised this empirical question, we see that also other senses of "free" may be relevant, for the attitudes of the wrongdoer may be such that he is more likely to be modified if he experiences the blame as justified than if he experiences it as unjustified, and perhaps he will not experience it as justified unless he acted freely in the following sense. . . . Hence, in the same sense in which it, on Nowell-Smith's view, may be an empirical requirement that a wrongdoer could have acted otherwise, it may also be an empirical requirement that he acted freely in other senses.

J. L. Austin

A lively debate over the correctness, especially the linguistic, but to some extent also the moral adequacy of interpreting "*P* could have acted otherwise" as "*P* would have acted otherwise, if he had chosen or decided or tried to" (T_2), has taken place within the decade, especially as a result of Austin's paper (1956).[38] Mild determinists have been attacked by a united front of indeterminists, libertarians, and hard determinists. The main lines of attack may be summarized as follows:

(*i*) T_2 does not express what we in ordinary language mean when we say that someone could have acted otherwise.

---

[37] Cp. *Hart* (1962), p. 25.

[38] The discussion elicited by Austin's paper includes contributions by *Lehrer* (1959–60), *Nowell-Smith* (1960), *O'Connor* (1960), *Taylor* (1960), *Goldberg & Heidelberger* (1960–61), *Lehrer* (1961–62), *Thalberg* (1961–62), *Bradley* (1962), *Kaufman* (1962), *Locke* (1962), *Dore* (1962–63), *Scarrow* (1962–63), *Whiteley* (1962–63), *Aune* (1962–63), *Baier* (1963), *Henschen-Dahlquist* (1963), *Dore* (1963–64), *Ewing* (1963–64), and *Honoré* (1964).

(*ii*) Freedom in the sense formulated in T_2 is not equivalent to freedom in the sense of having it in one's power to act otherwise.

(*iii*) Freedom in sense T_2 is not equivalent to freedom in the sense required by moral guilt and blame.

One of Austin's central contentions is that we use such sentences as "Smith could have acted otherwise" with indeterministic implications, and that if-analyses therefore are inconsistent with actual usage.

I see no reason to deny that so-called ordinary usage is flexible enough to include a usage of the kind indicated by Austin. But the question is whether this observation has any consequences for the free-will problem. It seems to me that it has no such implications. Even if everyone always used such sentences as "Smith could have acted otherwise" in an indeterministic sense, this would not imply that it is freedom in this sense which is required by moral responsibility. To find the morally relevant senses of "freedom," we must raise the question of freedom in relation to different types of moral systems, e.g., teleological and deontological systems. If we, e.g., presuppose act-utilitarianism, it certainly is not freedom in an indeterministic sense which is relevant.

The most fruitful part of Austin's paper is his attempt to show that such sentences as "Smith could have acted otherwise" express categorical statements about the past, they say that Smith was able to do something, and not that he would have been able to do something, if this or that condition had been fulfilled.

I consider it very likely that such sentences at least sometimes are used in this categorical way. But Austin's great mistake is to believe that this categorical sense must be an indeterministic sense, whereas it quite often refers to the agent's power to decide or act otherwise. But the crucial point, of course, is that linguistic hypotheses of this kind are not relevant to the free-will issue, for even if we assume that "could"-sentences always express categorical statements about our power to decide and act, this would neither imply that we have such power, nor that this sense of "freedom" is morally relevant. An act-utilitarian, e.g., will say that this sense is of no crucial importance to him, since all he wants to know is whether human beings are modifiable through the use of praise and blame. On the other hand, a philosopher accepting that human beings can be morally guilty may maintain that guilt presupposes that the agent had it in his power, in a categorical sense, to have acted otherwise, but this philosopher certainly would have to stand for his requirement even if no one actually used "could"-sentences so as to correspond with it.

Moreover, even if all of us always used "could"-sentences to express categorical statements, this does not imply that it would be incorrect to analyze "could"-sentences partly by the use of "would-have-if"-sentences.[39] Finally, even if it were completely misleading, from a linguistic point of view, to interpret "could"-sentences by the use of "would-have-if"-sentences, it might nevertheless be freedom in the sense of whether or not Smith would have acted otherwise if he had tried to, which is relevant from the point of view of moral guilt and blame. Many of the articles about how to analyze "can"- and "could"- sentences are uninteresting from the point of view of the free-will debate because the authors fail to see that the free-will problem partly is of moral nature.

The Role of the Self

To declare Smith morally guilty, we require that the wrong act, A_1, in some sense can be attributed to him.[40] Furthermore, he must have had it in his power to act otherwise, i.e., to do A_2 instead of A_1, and if A_2 had occurred, we require that it was *his* action. How shall we conceive of the relation between an action and the self?

Taylor (1963), and in his article in *Hook* (1958), *Campbell* (1957) and *Chisholm* (1964), among others, seem to postulate the existence of a transempirical or substantial self. Forgetting about the difficulties of interpreting the meaning of this use of "self" or what kind of evidence would be relevant for and against it, it is hard to understand the nature of the relation between the substantial or transempirical self and such empirical things as deciding and acting. As long as this problem has not been clarified, it seems reasonable to remain sceptical toward this way of connecting action and self.

---

[39] See, e.g., *Nowell-Smith* (1960), p. 93, and cp. also *Locke* (1962), *Whiteley* (1962–63), and *Ewing* (1963–64).

[40] Cp. in this connection *Wilson* (1958). See also Ralls' discussion of the relation between responsibility and personal identity (1963). Philippa Foot defends the tough line that "The argument that it will be a different *me* who will be beaten tomorrow carries no weight, for 'different' or not, the back which will be beaten is the one about which I am concerned today," (1962), p. 78.

R. B. Edwards and D. H. Whittier, defending an empirical notion of the self, have argued that "the self *is* its activities as well as its experiences" (see e.g., *Edwards* [1965], p. 274). If we accept this line of thinking, Smith's trying, deliberating, choosing, and deciding are activities of his self, and it may be empirically possible to connect his act to some of these activities and thereby to his self. But perhaps the relation between such activities and the self here becomes too intimate? It may be useful to distinguish, e.g., between "trying to do A occurs in Smith" and "Smith tries to do A." In order that the trying shall be an activity of his self, we must require, it seems to me, that it also in some ways corresponds to his norm- and value-system. But I shall not go into this question here, because, from the point of view of moral responsibility, the decisive thing is that the agent had control over those activities which led to his action, not that he considered them as parts of his self. An example: Suppose we had asked Eichmann whether or not he had it in his power to refrain from sending the Jews to the gas-chambers, and that he had answered: "Yes, if I had tried to stop the transportations, I would have succeeded, but such a trying would not have been a trying of the good and obedient, but of the bad and disobedient Eichmann, and hence it would not really have been my trying at all." This would not be a relevant excuse. Eichmann might be guilty even if the trying would not have been *his* trying in this sense of the term.

Let us turn to another point pertaining to the relation between action and self. Let us grant that Smith's deliberation was part of his self. How can we know that his action was brought about through his deliberation and not through something else, e.g., a certain neuro-physiological activity or a certain electric impulse from outside? Let us accept, for the sake of discussion, that Smith's deliberation, conceived as a mental or psychological process, in some way corresponds to a certain neuro-physiological process. Even if we then assume that this neuro-physiological process was caused by some previous neuro-physiological processes, and related by a law of nature to his action, this does not imply that his mental deliberation was unrelated to his action. To prove that it was, one would have to prove that the correlations between the act and the neuro-physiological processes in question would remain the same no matter what the characteristics of Smith's mental deliberation. And no such proof has been given.

Suppose it was not Smith's deliberation, but certain electric impulses transmitted to his brain, which led to A_1 and which would have led to A_2. How would this affect his responsibility? Whether or not we can attribute the action to Smith will depend upon his amount of control over these electric impulses. For instance, if he knew that if he decided not to kill Olga, then electric impulses would be transmitted to him in such a way that he would succeed in refraining from killing her, then he cannot excuse himself by saying that he did not have it in his power to refrain from killing her. He can excuse himself only if he had decisive reasons to believe that he was so completely powerless that there was no point in trying, and then he can excuse himself even if he actually would have succeeded in doing A_2, if he had tried to.

Power to Do What?

The requirement that the agent, to be responsible, must have had it in his power to act otherwise, is not satisfied by the fact that the agent had it in his power to do something else than what he did. A_2 must be different from A_1 in a morally relevant respect. To console Smith, driven by irresistible impulses toward killing Olga with his red gun, that he is free to kill her with his black gun, would be like saying to a man whose body has been trapped under a tree: "Why do you complain? You can still move your ears" (*O'Shaugnessy* [1956]). But even if his only alternative was to poison her, he may be guilty of having shot her instead of having poisoned her—even if he is not morally responsible for having killed her.

To say that the agent must have had it in his power to act otherwise does not imply that he must have had it in his power to succeed in doing that which he had it in his power to do instead of the thing he did. Smith is responsible for killing Olga even if the only thing he could have done instead would have been to *try* to comfort her, for trying to comfort her is succeeding in doing something else than killing her.[41]

Do We Have It in Our Power to Try?

Let us accept that "P, in S, had it in his power to do A_2 instead of A_1" implies "P, in S, would have done A_2 instead of A_1, if he had chosen or decided or tried to do A_2." (Let us select "if he had tried to do A_2" as representative of such conditions.) The latter sentence, however, does not imply that P had it in his power to try to do A_2. Suppose he did not have it in his power to try, in what sense did he then have it in his power to do A_2? Apparently only in the sense in which a beheaded person had it in his power to talk: "he would have talked if he had kept his head on." This is a serious objection to the would-have-if-analyses. The suggested solutions are:

---

[41] Cp. *Baier* (1963).

(*i*) "P had it in his power to try" must itself be analyzed by the use of an if-analysis, e.g., as "P would have tried, if he had tried to try or decided to try."

(*ii*) "P had it in his power to try" must be analyzed indeterministically, viz., as "P's trying or not-trying to do A_2 in S are uncaused."

(*iii*) It is misleading or meaningless to ask such questions as "Does he have it in his power to try?" "Does he have it in his power to choose?"

(*iv*) We always have it in our power to try.

(*v*) "P had it in his power to try" can be analyzed by the use of a categorical statement.

(*i*) does not work, because it leads to an endless regress. (*ii*) does not work because saying that Smith's trying was not caused does not imply that he had it in his power to try. Both things have often been pointed out, and I shall say nothing more about them.

Thalberg (1964) argues that such concepts as capacity, being able to, and probably also—although he does not say so explicitly—having something in one's power, have no application to normal cases of deciding and trying. Such concepts apply to actions, not to the will.[42] He tries to show this by an example: ". . . if we take a high school graduate who is faced with a choice of two or three universities, it would be puzzling to ask, "Can he make up his mind—has he the patience, the skill, the means for reaching a decision?" (p. 413). I am unable to find the question puzzling.

Let us turn to the fourth suggestion. According to Hampshire, Smith can excuse himself by saying that he tried with all his power and failed, or that he abstained from trying because he knew that he would fail (1959, p. 183), but he cannot excuse himself by saying that he did not have it in his power to try, for "there is no sense in which it is impossible for him to do something, provided that he knows what would be involved in doing it. He can always set out on the course of action, even if he knows that he will encounter difficulties at the very first stage of effective action, and even if he believes that he will in fact fail before any effective action has been taken" (pp. 182–83).

Whether or not we shall accept the thesis that in normal situations one can always try, depends upon what we mean by "trying." If we are willing to say that Smith tried not to kill Olga in so far as he whispered certain sounds to himself, then we may probably assert, on empirical grounds, that we always have it in our power to try,[43] for those who cannot whisper certain sounds to themselves are not rational animals. But if we take "trying" in a stronger sense referring to the making of an effort, perhaps even a strong and energetic effort like thinking about the consequences, studying the case quite closely, etc., then it is probably false that we can always try.

Those defending that we always have it in our power to try tend to interpret "try" as referring to what we can call "low-integrated" activities, where trying becomes something like talking to oneself. But thereby they decrease the likelihood that the statement "If he had tried, he would have done it" will be true unless in very unimportant cases, like taking one tie instead of another.

Suppose Smith had it in his power to try not to kill Olga, but that he did not think about the possibility of trying to do so. Let us even suppose that given his intelligence, upbringing, and imagination, he could not even be expected to think about this. Hampshire maintains that a person, in such a situation, can excuse himself (p. 186), and I agree with him, if "could not be expected to" implies that the person was powerless. It is a virtue of Hampshire's exposition that he discusses such cases and tries to combine limitations of power with limitations of intellect. But if he really wanted to develop a philosophy of the freedom of the mind, not only of the will, he should also have discussed the influence of a person's moral convictions. Suppose Smith would have done A_2 if he had tried to, suppose he did think about trying and that he had it in his power to try, but was firmly convinced that it would be morally wrong to try. Did he then have freedom of mind? A person's moral convictions may be one of the most important limitations of his power to try—for bad or for good. I may be aware of the possibility of killing my child. It may be true that if I try to do it, I shall do it, and I may know that this is so, but nevertheless be unable to do it, because I am unable to set out on a course of action which I so strongly feel to be morally wrong.

---

[42] Cp. also Pap, *op. cit.*

[43] *Skinner* (1963), defending the view that we have it in our power to try and to decide, seems to take "try" and "decide" in very weak senses.

Let us suppose that Smith did think about the possibility of trying not to kill Olga, that he did not consider it morally outrageous to let her live, and that he was not subject to an irresistible urge to kill. Given these suppositions, we conclude that Smith, so far as we can see, had it in his power to try, even in a strong sense of "try." But what can this mean, if we reject indeterministic or hypothetical analyses?

One possible answer is given by *Locke* (1962). He suggests that "*P* could have tried" means "*P* had the opportunity to try" which means that nothing prevented him from trying (p. 253). Locke's suggestion is valuable in so far as it interprets "*P* could try" as an empirical statement about *P*, but is, in the end, not very helpful. A person who is empty inside is not prevented from trying to do A_2, but does not for that reason have it in his power to try. On the other hand, a person may have it in his power to try even if his attempts will meet with strong inner resistance.

A number of authors have tried to show that to have something in one's power cannot always mean that one by doing so and so brings about this or that.[44] There must be something which one has in one's power in a more direct sense. If we had things in our power only in the former sense, we would have nothing in our power, because this sense, as indicated above, leads to an endless regress. Danto's concept of a basic action (1964), though not very clear, seems to be intended to delimit those activities which one has in one's power in a direct sense. According to his view, trying is a basic action; it is not something we bring about by first doing something else. But Danto seems to believe that if a certain activity, e.g., trying, is basic, then one can never bring it about, and this seems to me false. Any kind of activity, trying, deciding, thinking, etc., may be an example both of basic and not-basic activity. Even if trying usually is something that one does not bring about, one *may* bring it about, e.g., by deciding to try.

Danto tries to define a normal person in terms of a certain repertoire of basic actions, saying that persons who are unable to do these actions are abnormal. If we assume that trying is a basic action, this gives a nice "solution" of our problem: unless a person is abnormal, he is always able to try. The difficult task of finding out whether or not Smith had it in his power to try is reduced to the simpler task of finding out whether or not he was crazy. But this is unacceptable, for a person who in the strong sense of "try" could try to do any action, far from being normal, would be a very strange creature.

Even if we assume that a normal person can try to do any action in a weak sense of "try," and at least some actions in a stronger sense of "try," it does not follow that a particular person, e.g., Smith, had it in his power to try in a sense of "try" strong enough to make it likely that if he had tried, he would have done A_2 instead of A_1.

The above discussion suggests that the difficulty of formulating a categorical sense of "power" has nothing to do with the principle of causality, but is a result of our insufficient knowledge of the processes involved in trying, deciding, and acting. But no matter its wording, the relevant sense will make "*P* had it in his power to try" to an empirical statement, for to assert that a person had it in his power to try is to advance an empirical hypothesis about him. We may be able to show that the statement is false, because we can show that the person was prevented in some way. But even when everything looks perfect, and we are inclined to conclude that he had the power in question, we must reckon with the possibility that something did prevent him. Since such assertions are always uncertain in this respect, it seems a wise policy to admit that also our assertions about people's moral guilt do not rest on rocks.

Causality and Power

If the principle of causality applies to human behavior, e.g., to our decisions and actions, can we then have it in our power to decide or act otherwise than we in fact do? Given the two statements:

(*ii*) Smith's killing of Olga (A_1) in *S* was caused, and

(*iii*) Smith, in *S*, had it in his power to do A_2 instead of A_1. Does (*ii*) imply that (*iii*) is false?

(*iii*), we assume, can be taken as equivalent to

(*iv*) Smith, in *S*, would have done A_2 instead of A_1 if he had tried to, and he had it in his power to try to do A_2.

---

[44] See, e.g., *Pears* (1963), p. 24.

Obviously (*ii*) does not imply that (*iv*)'s first conjunct is false. The important question seems to be whether (*ii*) implies the falsity of

(*v*) Smith, in *S*, had it in his power to try to do A_2.

To get the problem as explicit as possible, we shall assume that Smith did not try to do A_2, and that his not-trying was caused. Let us reformulate (*ii*) thus: Smith, in *S*, killed Olga (A_1), and did not try to do A_2 (NTA_2), and both A_1 and NTA_2 were caused.

And for the purpose of discussing the relation to (*v*), we can shorten this reformulation:

(*i*) Smith's NTA_2 in *S* was caused.

The standard argument for showing that (*i*) implies that (*v*) is false, can be formulated thus: If NTA_2 was caused, then certain conditions were causally sufficient for NTA_2. If certain conditions were causally sufficient for NTA_2, then it was causally impossible for Smith, in *S*, to try to do A_2. If it was causally impossible for Smith to try to do A_2, then he could not try to do A_2. And if he could not try to do A_2, then he did not have it in his power to try to do A_2.[45]

An interesting feature of the recent free-will debate is the attempt to show that causality is compatible with power even if "power" is not analyzed in hypothetical terms. Let us first look at an article by *Canfield* (1962). Applied to our example, the main lines of his reasoning seem to be as follows:

(*a*) let us assume that Smith, in *S* (or, let us say, at time t_2), when he killed Olga, had the ability to refrain from killing her as well as the ability to kill her.

Let us refer to the state of Smith at t_2 as "State Two," and by "State One" to that state, whatever it may be, whose description is a detailed description of Smith and his environment at some time prior to t_2: t_1. Let us now formulate the true statement:

(*b*) At t_1 Smith was in State One and at t_2 in State Two. Let us also presuppose that the following general statement is true:

(*c*) Whenever someone is in State One at time t_n, at time t_n+1 he will be in State Two.

Since (*b*) is an instantiation of (*c*), (*c*) is consistent with (*b*). The important question, according to Canfield, is whether (*b*) is consistent with (*c*) even if we assume that (*c*) is not only true, but a so-called law of nature, e.g., universal in form, non-tautologous, convertible to counter-factual, and simple (see, p. 364). He concludes that it is, and that it therefore is logically possible to give a causal explanation of being in State Two, which implies that it is logically possible to hold both that Smith's not-trying to refrain from killing Olga was caused and that he at the same time had the ability to try to refrain from killing her.

Canfield tries to explain what he means by "free will," but says very little about what he means by "ability," and does not try to show that his interpretation of this term (whatever it is) is relevant to questions of guilt and blame. He illustrates what he means by the example of a person who is tempted to laugh, but keeps himself from laughing. This person, he says, does not only have the ability to refrain from laughing, but also the ability to laugh.

I find this view very strange. Applied to my own example, it implies that from the statement that Smith refrained from letting Olga live, we can infer that he was able (or even: had the ability) to let her live! It is this strange thinking about abilities which is the really weak point in Canfield's article.

In his answer to Dore's criticism of his article (1963),[46] *Canfield* (1963) argues as if he means that saying that a person has an ability to do something is to make a categorical, empirical statement about that person, for he mentions that a physical examination might tell us whether or not a certain person had certain abilities. But the important thing is to develop and defend this interpretation of "ability" (or better: of "was able to"). Once that has been done, the assertion that causality is compatible with ability in such a sense is not more astonishing than the assertion that causality is compatible with being healthy.

Another, and perhaps more promising line of thinking is found in an article by *Aune* (1962–63). He follows up Moore's well-known suggestion that we must distinguish between two different senses of "could-not," and develops it in a fruitful way.

Since NTA_2 was caused, certain conditions, say *C*, were causally sufficient for NTA_2. From this it follows that

(*a*) it would be inconsistent with certain laws of nature to assert that *C* was the case and NTA_2 did not occur.

This inference is not uncontroversial, for it is relative to a certain interpretation of "was caused" and makes use of such expressions as "laws of nature," but we shall assume, for the sake of argument, that the inference must be accepted.

---

[45] See, e.g., *Taylor* (1960), p. 89; and cp. *Ayer* (1963), p. 265.

[46] Canfield's arguments are criticized also by *Lehrer* (1964.1).

We can formulate (*a*) also in the following way:

(*b*) It could not have been the case that C occurred and Smith, in S, tried to do A_2.

According to Aune, it is important to see that (*b*) uses "could-not" as referring to a conjunction of factors. It says that certain conditions and a certain trying to do A_2 "add up to an impossible totality, a totality that could not have existed" (p. 411). The "could-not" does not say or imply that there was something Smith could not do. Those maintaining that (*i*) implies that (*v*) is false confuses two different senses of "could-not," one referring to the total lack of likelihood that a certain system, *Smith-in-S*, shall give a certain output (an attempt to do A_2), the other referring to the powerlessness of a certain person. (*i*) implies the first sense of "could not". (*v*) denies "could not" in the second sense.[47]

Power and Empirical Conditions

The fact that the principle of causality does not annihilate or reduce our power to decide or act otherwise than we in fact do, does not imply that we have it in our power to decide and act otherwise. Our power may be limited by lack of imagination and knowledge, inner conflicts and ethical convictions, etc. Hence, what one man can do is not necessarily what another can do. What we can do in one situation may be different from what we can accomplish in another. And even if our power to make a certain well-integrated decision *now* is quite limited, it may increase through patient endeavor.

Since our power to decide and act is related to empirical conditions, it may be fruitful to ask what kind of variations in these conditions will lead to this or that degree of power, or what kind of strategies a person may make use of to increase his power in a certain respect?—Very few philosophers have dealt with questions of this kind.[48]

Ethics and Freedom as Power

The thesis that an agent cannot be morally guilty if he was powerless to act otherwise is probably accepted by quite a few philosophers, but it can be denied without making logical or empirical errors, for it is a normative standpoint as regards one of the conditions of moral guilt. As long as we have not shown that this normative standpoint is more correct than alternative standpoints, the free-will problem is not solved. A philosopher might hold an optimistic opinion concerning our power to decide and act otherwise and nevertheless maintain that we have no free will, because he holds that moral guilt does not presuppose freedom in this sense, but in a sense in which we are not free. On the other hand, a philosopher maintaining that we do not have it in our power to decide and act otherwise, may nevertheless hold that we have a free will because he, e.g., believes that the notion of moral guilt is nonsensical, and that the question of moral responsibility is the question of the usefulness of blame. If he could show that this is the morally correct concept of responsibility, the positions of the indeterminist, the libertarian, and the hard determinist or evolutionist would be without interest to the moral philosopher.

Austin's "A Plea for Excuses" (1956–57) was one of the first publications within the recent debate to stimulate interest in the moral context of the free-will problem, but judging from his later "Ifs and Cans," he did not clearly understand the implications of his own approach.

Beardsley (1960–61) attempts to show that judgments of moral praise and blame may be made from what she called different "moral perspectives," e.g., the perspectives of "moral worth," "moral credit," and of "ultimate moral equality." She also tried to show that the sense of "freedom" required by praise and blame depends upon the perspective from which the judgment is made. This is a fruitful insight, but the author should have related the different perspectives to different types of moral systems, and analyzed the problems without assuming that judgments of praise and blame can be true or false. Because she does not do this, she sees the three perspectives as supplementing each other, and does not pay enough attention to the conflicts among them.

The moral relativity of the free-will problem is a central theme in *Ofstad* (1961), and it is stressed, among others, by *Kaufman* (1962), *Whiteley* (1962–63), and *Bronaugh* (1963–64).

---

[47] Cp. also *Ofstad* (1961), pp. 91, 241. *Hampshire*, in his latest book (1965), seems to take a similar position. He says: " 'There is no possibility of his acting differently' does not entail 'He cannot act differently' " (p. 32).

[48] But see *Hampshire* (1960), e.g., pp. 190 ff. and *Ofstad* (1961), pp. 185–203.

Moral Guilt and Freedom

Power to act otherwise is not the only condition of moral guilt. Knowledge of facts and the acceptance of moral norms are necessary, too, but it is the only condition which interests us in this article.

Moral guilt, just as criminal guilt, should be conceived of in such a way that it can vary in degrees.[49] An agent may be guilty to a high degree if he intentionally did evil and had a high degree of power to act otherwise. He may be guilty to a low degree if he either did not expect any harmful consequences of his behavior or had a very low degree of power to act otherwise. A maximum of guilt presupposes power to act otherwise in its highest possible degree: no effort at all was needed to act rightly. Seldom, if ever, do we have this amount of power to act otherwise, especially not in situations where we acted wrongly. It may be emotionally convenient to believe that the great evildoers are maximally guilty, but quite often they seem to be closer to evil things than evil agents, petrified by their limited power to act otherwise as well as by their limited knowledge and moral understanding.

Does moral guilt presuppose power to decide otherwise, not only power to act otherwise?

I have already taken a standpoint on this issue, for I have interpreted "P, in S, had it in his power to do A_2 instead of A_1," not as "P, in S, would have done A_2 instead of A_1 if he had chosen or decided or tried to do so," but as "P in S, would have done A_2 instead of A_1 if he had chosen or decided or tried to do so, and he had it in his power to choose or decide or try to do so." Hence, power to fulfill the if-clause is a condition of moral guilt. Moreover, a person may be responsible for his choices, decisions, and attempts. And in those cases his power at the action-level is irrelevant.

Power to decide and act otherwise is a condition of moral guilt because a wrongdoer who was powerless to act rightly is not a morally bad man. An agent who acted wrongly in spite of the fact that he tried with all his energy to act rightly is morally a good person. So is also a wrongdoer who on sufficiently good grounds held the false opinion that he was so completely powerless to act rightly that he had no reason to try. But what should we say about a wrongdoer who did not have it in his power to act rightly, but believed that he did, nevertheless, refraining from even trying to do so? It seems to me that such a person may be just as morally bad as one who actually had the power in question. Hence, power to act otherwise is not a necessary condition of moral guilt, but the absence of such power may be a sufficient condition of moral innocence.

Moral Blame and Freedom

Just as we may have to catch a tiger and lock it up, we may have to catch a man running wild. And just as we may blame a dog, we may blame a man just because it is useful. But in such cases we should talk about non-moral, rather than moral, blame.

Moral blame presupposes moral guilt, but guilt is not a sufficient condition. To blame a person is to act toward him in a certain way, and an act is unjustified if it makes the world worse than if we acted otherwise. Hence, moral blame presupposes guilt as well as utility.[50] And we must ask: Is blame a useful instrument of reformation, prevention, deterrence, or moral education? We do not know, but a large number of studies on the effects of punishment suggest that the use of blame is a very ineffective way of modifying people's behavior.[51] Perhaps blaming primarily fulfills a function for the moral judge, whereas argumentation, knowledge, or a nicer wife are more effective instruments of change. Hence, though it is not logically contradictory to assert that a wrongdoer should be morally blamed, the conditions of moral blame: guilt and utility, are seldom, if ever, fulfilled, at least not to a high degree.

X. CONCLUSION

Let me conclude with a short survey of the positive and negative traits of the different positions.

The indeterminists' notion of freedom is irrelevant from the point of view of moral guilt and blame.

The libertarians have made a contribution through their emphasis on such factors as trying, deciding, making an effort, and through their insight that moral responsibility, at least in one sense, presupposes the freedom of the agent. But by conceiving of this agent as a transempirical substance, they have been led on to strange and unclear theories of "doing," and have found it necessary to defend indeterminism to try to safeguard the possibility of the operation of such a transempirical substance.

---

[49] Cp. *Hart* (1962), p. 17.

[50] For a similar point of view applied to legal responsibility, see *Hart* (1962), p. 30.

[51] See, e.g., W. K. Estes "An Experimental Study of Punishment," *Psychological Monographs*, vol. 57 (1944).

Both the hard determinists and the libertarians are right in insisting that the question of moral responsibility is not only the question of the usefulness of praise or blame, but also of attributing right or wrong actions to agents. Hence, they are right in denying that the free-will issue can be reduced to the question of the modifiability of behavior, and in asserting that responsibility presupposes freedom in a categorical sense. But by failing to see either that "freedom" here must be interpreted as power to decide and act otherwise, or that such power is compatible with the principle of causality, they have arrived at the conclusion that since the principle of causality must be accepted, we are not free in the sense required by moral responsibility.

The hard evolutionists understand, or ought to understand, that our freedom is not threatened by the principle of causality, but by certain types of causal explanations, but they have failed to make use of this insight in order to explore the empirical conditions of varying degrees of freedom.

The mild determinists have tended to neglect responsibility in the senses discussed by the libertarians and the hard determinists and evolutionists, and they have, just as these, paid insufficient attention to the notion of moral guilt. They are right in maintaining that modifiability is a necessary condition of moral blame, but it is not a sufficient condition, and it is neither necessary nor sufficient for moral guilt.

* * *

Solving the free-will problem implies solving three different kinds of problems: (*i*) In what senses of "moral responsibility," if any, is it correct to assert that we may be morally responsible? (*ii*) In what senses of "freedom," if any, is freedom a condition of moral responsibility as clarified in (*i*)? (*iii*) Under what conditions, if any, are we free in the senses clarified in (*ii*)? Hence, the solution involves the clarification of moral, analytical, and scientific problems. The recent debate has been concerned mostly with the analytical problems and has been too optimistic about the possibility that linguistic analysis would lead to the solution of the whole problem. Fruitful research presupposes that the analytical instruments are used within a well-defined moral context and in close contact with scientific research on man's behavior.

University of Stockholm

Received May 25, 1966

BIBLIOGRAPHY*

ADLER, M. J. (ed.) *The Idea of Freedom*, 2 vols. (New York; 1958, 1961).

ANSCOMBE, G. E. M. *Intention* (Oxford, 1957).

——— "The Two Kinds of Error in Action" (Symposium: Human Action), *The Journal of Philosophy*, vol. 60 (1963), pp. 393–401.

AUNE, B. "Abilities, Modalities, and Free Will," *Philosophy and Phenomenological Research*, vol. 23 (1962–1963), pp. 397–413.

AUSTIN, J. L. "Ifs and Cans," *Proceedings of the British Academy*, London, vol. 42 (1956), pp. 109–132 (reprinted in *Philosophical Papers*, [Oxford, 1961]).

——— "A Plea for Excuses," *Proceedings of the Aristotelian Society*, vol. 57 (1956–1957), pp. 1–30 (reprinted in *Philosophical Papers*, [Oxford, 1961]).

AYER, A. J. "Fatalism" in *The Concept of a Person* (London, 1963).

BAIER, K. "Could and Would," *Analysis Supplement*, vol. 23 (1963), pp. 20–29.

——— "Action and Agent," *The Monist*, vol. 49 (1965), pp. 183–195.

BAYLIS, C. A. *Ethics* (New York, 1958), Ch. 2.

---

* This bibliography covers works written in English and printed 1954–1965. Articles have been checked in the following periodicals: *American Philosophical Quarterly, Analysis, The Australasian Journal of Philosophy, The British Journal for the Philosophy of Science, Ethics, The Journal of Philosophy, Mind, The Monist, Proceedings of the British Academy, The Philosophical Quarterly, The Philosophical Review, Philosophy, Philosophy and Phenomenological Research, Proceedings of the Aristotelian Society* (including supplementary volumes), *Theoria*, and *Inquiry*. Some minor articles and notes have been left out.

Useful bibliographies will be found in *Adler* (1958/61), and *Ofstad* (1961). See also the short bibliography by Roland Hall in *The Philosophical Quarterly*, vol. 15 (1965), pp. 179–181, and especially the annotated bibliography in P. Edwards and A. Pap (eds.), *A Modern Introduction to Philosophy* (revised edition, New York, 1965).

BEARDSLEY, E. L. "Determinism and Moral Perspectives," *Philosophy and Phenomenological Research*, vol. 21 (1960–1961), pp. 1–20.

BECK, L. "Agent, Actor, Spectator, and Critic," *The Monist*, vol. 49 (1965), pp. 167–182.

BENNETT, J. "The Status of Determinism," *The British Journal for the Philosophy of Science*, vol. 14 (1963–1964), pp. 106–119.

BERLIN, I. "From Hope and Fear Set Free," *Proceedings of the Aristotelian Society*, vol. 64 (1963–1964), pp. 1–30.

BEROFSKY, B. "Determinism and the Concept of a Person," *The Journal of Philosophy*, vol. 61 (1964), pp. 461–475.

BOHM, D. *Causality and Chance in Modern Physics* (London, 1957).

BORING, E. G. "When Is Human Behavior Predetermined?" in D. E. Dulany (ed.), *Contributions to Modern Psychology* (Oxford, 1963).

BRANDT, R. "Blameworthiness and Obligation," in A. I. Melden (ed.), *Essays in Moral Philosophy* (Seattle, 1958).

——— *Ethical Theory* (Englewood Cliffs, 1959), Ch. XX.

——— and KIM, J. "Wants as Explanations of Action" (Symposium: Human Action), *The Journal of Philosophy*, vol. 60 (1963), pp. 425–435.

BRADLEY, M. C. "A Note on Mr. MacIntyre's *Determinism*," *Mind*, vol. 68 (1959), pp. 521–526.

BRADLEY, R. D. "Free Will: Problem or Pseudo-problem?" *The Australasian Journal of Philosophy*, vol. 36 (1958), pp. 33–45.

——— " 'Ifs,' 'Cans,' and Determinism," *The Australasian Journal of Philosophy*, vol. 40 (1962), pp. 146–158.

——— "Determinism or Indeterminism in Microphysics," *The British Journal for the Philosophy of Science*, vol. 13 (1962–1963), pp. 193–215.

——— "Causality, Fatalism, and Morality," *Mind*, vol. 72 (1963), pp. 591–594.

BRAITHWAITE, R. B. (ed.) "*Analysis* Problem No. 7," *Analysis*, vol. 16 (1955), pp. 1–5. (Contributions by Brian Ellis, "Candidus," and Nicholas Rescher.)

BRETHERTON, D. " 'Ought' Implies 'Can Say'," *Proceedings of the Aristotelian Society*, vol. 63 (1962–1963), pp. 145–166.

BRONAUGH, R. "Freedom as the Absence of an Excuse," *Ethics*, vol. 74 (1964), pp. 161–173.

BUNGE, M. *Causality: The Place of the Causal Principle in Modern Science* (Cambridge, 1959).

CAMPBELL, C. A. *On Selfhood and Godhood* (London, 1957), Lecture IX.

——— "Free Will: A Reply to Mr. R. D. Bradley," *The Australasian Journal of Philosophy*, vol. 36 (1958), pp. 46–56.

——— "Moral Libertarianism: A Reply to Mr. Franklin," *The Philosophical Quarterly*, vol. 12 (1962), pp. 337–347.

CANFIELD, J. "Determinism, Free-Will and the Ace Predictor," *Mind*, vol. 70 (1961), pp. 412–416.

——— "Knowing about Future Decisions," *Analysis*, vol. 22 (1961–1962), pp. 127–129.

——— "The Compatibility of Free Will and Determinism," *The Philosophical Review*, vol. 71 (1962), pp. 352–368.

——— "Free Will and Determinism: A Reply," *The Philosophical Review*, vol. 72 (1963), pp. 502–504.

CHISHOLM, R. M. "The Descriptive Element in the Concept of Action" (Symposium: Action, Motive, and Desire), *The Journal of Philosophy*, vol. 61 (1964.1), pp. 613–625.

——— *Human Freedom and the Self*, The Lindley Lecture of the University of Kansas (Kansas, 1964).

COHEN, M. F. "Motives, Causal Necessity and Moral Accountability," *The Australasian Journal of Philosophy*, vol. 42 (1964), pp. 322–334.

DANTO, A. C. "The Paradigm Case Argument and the Free Will Problem," *Ethics*, vol. 69 (1958–1959), pp. 120–124.

——— "What We Can Do" (Symposium: Human Action), *The Journal of Philosophy*, vol. 60 (1963), pp. 435–445.

——— "Basic Actions," *American Philosophical Quarterly*, vol. 2 (1965), pp. 144–148.

——— and MORGENBESSER, S. "Character and Free Will," *The Journal of Philosophy*, vol. 54 (1957), pp. 493–505.

D'ARCY, E. *Human Acts* (Oxford, 1963), Ch. 3.

DAVENEY, T. F. "Choosing," *Mind*, vol. 73 (1964), pp. 515–526.

DAVIDSON, D. "Actions, Reasons, and Causes," *The Journal of Philosophy*, vol. 60 (1963), pp. 685–700.

DILMAN, I. "The Freedom of Man," *Proceedings of the Aristotelian Society*, vol. 62 (1961–1962), pp. 39–62.

DIODORUS CROMUS "Time, Truth and Ability," *Analysis*, vol. 26 (1965–1966), pp. 137–147.

DONNELLAN, K. S. "Knowing What I am Doing" (Symposium: Human Action), *The Journal of Philosophy*, vol. 60 (1963), pp. 401–409.

DORE, C. "On the Meaning of 'Could Have'," *Analysis*, vol. 23 (1962–1963), pp. 179–181.

——— "More on the Meaning of 'Could Have'," *Analysis*, vol. 24 (1963–1964), pp. 41–43.

——— "Is Free Will Compatible with Determinism?" *The Philosophical Review*, vol. 72 (1963), pp. 500–501.

EDWARDS, R. "Agency Without a Substantive Self," *The Monist*, vol. 49 (1965), pp. 273–289.

EVANS, J. L. "Choice," *The Philosophical Quarterly*, vol. 5 (1955), pp. 303–315.

EWING, A. C. *Second Thoughts in Moral Philosophy* (London, 1959).

——— "May Can-Statements be Analysed Deterministically?" *Proceedings of the Aristotelian Society*, vol. 64 (1963–1964), pp. 157–176.

FAIN, H. "Prediction and Constraint," *Mind*, vol. 67 (1958), pp. 366–378.

FALK, W. D. "Action-guiding Reasons" (Symposium: Action-guiding Reasons), *The Journal of Philosophy*, vol. 60 (1963), pp. 702–718.

FARRER, A. *The Freedom of the Will* (London, 1958).

FEINBERG, J. "On Being 'Morally Speaking' a Murderer," *The Journal of Philosophy*, vol. 61 (1964), pp. 158–171.

——— "Action and Responsibility" in Max Black (ed.), *Philosophy in America* (Ithaca, 1965).

FINGARETTE, H. "Psychoanalytic Perspectives on Moral Guilt and Responsibility: A Re-Evaluation," *Philosophy and Phenomenological Research*, vol. 16 (1955–1956), pp. 18–36.

FLEW, A. "Divine Omnipotence and Human Freedom" in A. Flew and A. MacIntyre (eds.), *New Essays in Philosophical Theology* (London, 1955).

——— "Determinism and Rational Behaviour," *Mind*, vol. 68 (1959), pp. 377–382.

FOOT, P. "Free Will as Involving Determinism," *The Philosophical Review*, vol. 66 (1957), pp. 439–450 (reprinted in Morgenbesser & Walsh, *Free Will* [1962]).

FRANKLIN, R. L. "Dissolving the Problem of Free-will," *The Australasian Journal of Philosophy*, vol. 39 (1961), pp. 111–124.

——— "Moral Libertarianism," *The Philosophical Quarterly* vol. 12 (1962), pp. 24–35.

FRIEDRICH, C. J. (ed.) "Responsibility" in *Nomos* III (New York, 1960).

GALE, R. M. "Endorsing Predictions," *The Philosophical Review*, vol. 70 (1961), pp. 376–385.

GALLAGHER, K. T. "On Choosing to Choose," *Mind*, vol. 73 (1964), pp. 480–495.

GALLIE, W. B. *Free Will and Determinism Yet Again*, An Inaugural lecture, (The Queens University of Belfast, 1957).

GALLOP, D. "On Being Determined," *Mind*, vol. 71 (1962), pp. 181–196.

GASKING, E. "Causation and Recipes," *Mind*, vol. 64 (1955), pp. 479–487.

GEACH, P. T. "Ascriptivism," *The Philosophical Review*, vol. 69 (1960), pp. 221–225.

GINET, C. "Can the Will be Caused?" *The Philosophical Review*, vol. 71 (1962), pp. 49–55.

GLASGOW, W. D. "On Choosing," *Analysis*, vol. 17 (1956–1957), pp. 135–139.

——— "The Concept of Choosing," *Analysis*, vol. 20 (1959–1960), pp. 63–67.

GLASS, B. "The Relation of the Physical Sciences to Biology-Indeterminacy and Causality" in B. Baumrin (ed.), *Philosophy of Science, The Delaware Seminar* (New York, 1963).

GOLDBERG, B. and HEIDELBERGER, H. "Mr. Lehrer on the Constitution of Cans," *Analysis*, vol. 21 (1960–1961), p. 96.

GRÜNBAUM, A. "Complementarity in Quantum Physics and its Philosophical Generalization" (Symposium: Determinism in the Light of Recent Physics), *The Journal of Philosophy*, vol. 54 (1957), pp. 713–727.

GUSTAFSON, D. F. "Voluntary and Involuntary," *Philosophy and Phenomenological Research*, vol. 24 (1963–1964), pp. 493–501.

HAINES, N. "Responsibility and Accountability," *Philosophy*, vol. 30 (1955), pp. 141–163.

HAKSAR, V. "Aristotle and the Punishment of Psychopaths," *Philosophy*, vol. 39 (1964), pp. 323–340.

HALVERSON, W. H. "The Bogy of Chance," *Mind*, vol. 73 (1964), pp. 567–570.

HAMLYN, D. W. "Causality and Human Behaviour," *Proceedings of the Aristotelian Society*, supplementary vol. 38 (1964), pp. 125–142.

HAMPSHIRE, S. and HART, H. L. A. "Decision, Intention and Certainty," *Mind*, vol. 67 (1958), pp. 1–12.

HAMPSHIRE, S. *Thought and Action* (London, 1959).

——— "Reply to Walsh on Thought and Action" (Symposium: Human Action), *The Journal of Philosophy*, vol. 60 (1963), pp. 410–424.

——— *Freedom of the Individual* (London, 1965).

HANDY, R. "Determinism, Responsibility, and the Social Setting," *Philosophy and Phenomenological Research*, vol. 20 (1959–1960), pp. 469–476.

HANSON, N. R. "Causal Chains," *Mind*, vol. 64 (1955), pp. 289–311.

HARDIE, W. F. R. "My Own Free Will," *Philosophy*, vol. 32 (1957), pp. 21–38.

HARE, R. M. *Freedom and Reason* (Oxford, 1963), Ch. 4.

HART, H. L. A. "Prolegomenon to the Principles of Punishment," *Proceedings of the Aristotelian Society*, vol. 60 (1959–1960), pp. 1–26.

——— "Negligence, *Mens Rea*, and Criminal Responsibility" in A. G. Guest (ed.), *Oxford Essays in Jurisprudence* (Oxford, 1961; reprinted in Morgenbesser and Walsh, *Free Will* [1962]).

——— *Punishment and the Elimination of Responsibility*, L. T.

——— Hobhouse Memorial Trust Lecture, University of London, 1962.

——— *see* Hampshire, S. and Hart, H. L. A. (1958).

HARTSHORNE, C. "Freedom Requires Indeterminism and Universal Causality," *The Journal of Philosophy*, vol. 55 (1958), pp. 793–811.

HEDENIUS, I. "Broad's Treatment of Determinism and Free Will" in P. A. Schlipp (ed.), *The Philosophy of C. D. Broad* (New York, 1959).

HEIDELBERGER, H. *see* Goldberg, B., and Heidelberger, H. (1960–1961).

HENSCHEN-DAHLQUIST, Ann-Mari, "Remarks to Austin's Criticism of Moore's Analysis of 'can'," *Theoria*, vol. 29 (1963), pp. 305–315.

HERBST, P. "Freedom and Prediction," *Mind*, vol. 66 (1957), pp. 1–27.

HONORÉ, A. M. "Can and Can't," *Mind*, vol. 73 (1964), pp. 463–479.

HOOK, S. (ed.) *Determinism and Freedom in the Age of Modern Science* (New York, 1958). (Contributions by: W. Barrett, E. L. Beardsley, M. Black, B. Blanshard, R. Brandt, P. W. Bridgman, R. M. Chisholm, C. J. Ducasse, P. Edwards, H. L. A. Hart, C. G. Hempel, H. W. Hintz, S. Hook, J. Hospers, A. Landé, A. Lerner, M. K. Munitz, E. Nagel, F. S. C. Northrop, A. Pap, A. Schultz, D. W. Sciama, R. Taylor, P. Weiss, and H. Wilson.)

HOSPERS, J. *Human Conduct* (New York, 1961),Ch. 10.

KAUFMAN, A. S. "Moral Responsibility and the Use of 'Could Have'," *The Philosophical Quarterly*, vol. 12 (1962), pp. 120–128.

——— "Ability," *The Journal of Philosophy*, vol. 60 (1963), pp. 537–551.

KENNER, L. "Causality, Determinism and Freedom of the Will," *Philosophy*, vol. 39 (1964), pp. 233–248.

KENNY, A. *Action, Emotion and Will* (London, 1963).

KIM, J. *see* Brandt, R. and Kim, J. (1963).

KÖRNER, S. "Science and Moral Responsibility," *Mind*, vol. 73 (1964), pp. 161–172.

LACEY, A. R. "Free Will and Responsibility," *Proceedings of the Aristotelian Society*, vol. 58 (1957–1958), pp. 15–32.

LEHRER, K. "Can We Know That We Have Free Will by Introspection?" *The Journal of Philosophy*, vol. 57 (1960), pp. 145–157.

——— "Ifs, Cans, and Causes," *Analysis*, vol. 20 (1959–1960), pp. 122–124.

——— "Cans and Conditionals: A Rejoinder," *Analysis*, vol. 22 (1961–1962), pp. 23–24.

——— "Decisions and Causes," *The Philosophical Review*, vol. 72 (1963), pp. 224–227.

——— " 'Could' and Determinism," *Analysis*, vol. 24 (1963–1964), pp. 159–160.

——— "Doing the Impossible," *The Australasian Journal of Philosophy*, vol. 42 (1964.1), pp. 86–97.

——— "Doing the Impossible; A Second Try," *The Australasian Journal of Philosophy*, vol. 42 (1964.2), pp. 249–251.

——— and TAYLOR, R. "Time, Truth, and Modalities," *Mind*, vol. 74 (1965), pp. 390–398.

LEWIS, H. D. "Freedom and Responsibility" in *idem*, *Freedom and History*, (London, 1962).

LOCKE, D. "Ifs and Cans Revisited," *Philosophy*, vol. 37 (1962), pp. 245–256.

LYON, A. "The Prediction Paradox," *Mind*, vol. 68 (1959), pp. 510–517.

MABBOTT, J. D. "Freewill and Punishment" in H. D. Lewis (ed.), *Contemporary British Philosophy* (London, 1956).

——— "Free Will" in *Encyclopaedia Britannica* (1957).

MACINTYRE, A. C. "Determinism," *Mind*, vol. 66 (1957), pp. 28–41.

MACKAY, D. M. "On the Logical Indeterminacy of a Free Choice," *Mind*, vol. 69 (1960), pp. 31–40.

MACKIE, J. L. "Responsibility and Language," *The Australasian Journal of Philosophy*, vol. 33 (1965), pp. 143–159.

——— "Evil and Omnipotence," *Mind*, vol. 64 (1955), pp. 200–212.

MACKINNON, D. M. *A Study in Ethical Theory* (New York, Collier Book, 1962), Ch. 4.

MACMURRAY, J. *The Self as Agent* (London, 1957).

MADDEN, E. H. "Psychoanalysis and Moral Judgeability," *Philosophy and Phenomenological Research*, vol. 18 (1957–1958), pp. 68–79.

MANDELBAUM, M. "Determinism and Moral Responsibility," *Ethics*, vol. 70 (1959–1960), pp. 204–219.

MATSON, W. I. "On the Irrelevance of Free-Will to Moral Responsibility, and the Vacuity of the Latter," *Mind*, vol. 65 (1956), pp. 489–497.

MAYO, B. A. "A Logical Limitation on Determinism," *Philosophy*, vol. 33 (1958.1), pp. 50–55.

——— *Ethics and the Moral Life* (London, 1958.2), Ch. XII.

——— "The Open Future," *Mind*, vol. 71 (1962), pp. 1–14.

McGILL, V. J. "Conflicting Theories of Freedom," *Philosophy and Phenomenological Research*, vol. 20 (1959–1960), pp. 437–452.

MELDEN, A. I. *Free Action* (London, 1961).

MILLER, G. "An Examination of Ofstad's Radical Contextualism," *Inquiry*, vol. 7 (1964), pp. 209–218.

MONTEFIORE, A. "Determinism and Causal Order," *Proceedings of the Aristotelian Society*, vol. 58 (1957–1958), pp. 125–142.

——— " 'Ought' and 'Can'," *The Philosophical Quarterly*, vol. 8 (1958), pp. 24–40.

MORGENBESSER, S., and WALSH, J. (eds.), *Free Will* (Englewood Cliffs, 1962). (See articles by H. L. A. Hart, P. Foot, and R. Taylor).

MORGENBESSER, S. *see* Danto, A. C., and Morgenbesser, S.

MORRIS, H. (ed.) *Freedom and Responsibility* (Stanford, 1961).

MUNK, A. "The Self as Agent and Spectator," *The Monist*, vol. 49 (1965), pp. 262–272.

MUNN, A. M. *Free-Will and Determinism* (London, 1960).

MURPHY, A. "Jonathan Edwards on Free Will and Moral Agency," *The Philosophical Review*, vol. 68 (1959), pp. 181–202.

NAGEL, E. *The Structure of Science* (New York, 1961), Ch. 10.

NOWELL-SMITH, P. H. *Ethics* (New York and London, 1954), Chs. 19–20.

——— "Determinists and Libertarians," *Mind*, vol. 63 (1954), pp. 317–337.

——— "Choosing, Deciding and Doing," *Analysis*, vol. 18 (1957–1958), pp. 63–69.

——— "Ifs and Cans," *Theoria*, vol. 26 (1960), pp. 85–101.

O'CONNOR, D. J. "Determinism and Predictability," *The British Journal for the Philosophy of Science*, vol. 7 (1956–1957), pp. 310–315.

——— "Possibility and Choice," *Proceedings of the Aristotelian Society, Supplementary Volume* 34 (1960), pp. 1–24.

OLDENQUIST, A. "Causes, Predictions, and Decisions," *Analysis*, vol. 24 (1963–1964), pp. 55–58.

OFSTAD, H. "Broad on Ought and Can," *Theoria*, vol. 21 (1955), pp. 105–116.

——— "Can We Produce Decisions?" *The Journal of Philosophy*, vol. 56 (1959), pp. 89–94.

——— "Frankena on Ought and Can," *Mind*, vol. 68 (1959), pp. 73–79.

——— *An Inquiry into the Freedom of Decision* (Oslo-Stockholm-London, 1961).

——— "Libertarianism and the Belief in Transempirical Entities, L. J. Russell on Causation and Agency" in *Philosophical Essays Dedicated to Gunnar Aspelin* (Lund, 1963).

OPPENHEIM, F. *Dimensions of Freedom* (New York, 1961).

O'SHAUGHNESSY, B. "The Limits of the Will," *The Philosophical Review*, vol. 65 (1956), pp. 443–490.

——— "Observation and the Will" (Symposium: Human Action), *The Journal of Philosophy*, vol. 60 (1963), pp. 367–392.

PAPANOUTSOS, E. P. "Freedom and Causality," *Philosophy*, vol. 34 (1959), pp. 193–203.

PEARS, D. F. (ed.) *Freedom and the Will* (London, 1963). (Contributions by: P. L. Gardiner, S. Hampshire, H. L. A. Hart, I. Murdoch, D. F. Pears, P. F. Strawson, J. F. Thomson, G. J. Warnock, M. Warnock, and B. Williams.)

PERRY, D. "Prediction, Explanation and Freedom," *The Monist*, vol. 49 (1965), pp. 234–247.

PETERS, R. S. *The Concept of Motivation* (London, 1958).

PIKE, N. "Divine Omniscience and Voluntary Action," *The Philosophical Review*, vol. 74 (1965), pp. 27–46.

PITCHER, G. "Hart on Action and Responsibility," *The Philosophical Review*, vol. 69 (1960), pp. 226–235.

——— Necessitarianism," *The Philosophical Quarterly*, vol. 11 (1961), pp. 201–212.

PLANTINGA, A. "The Free Will Defence" in Max Black (ed.), *Philosophy in America* (Ithaca, 1965).

PLAUT, H. C. "Condition, Cause, Free Will, and the Direction of Time," *The British Journal for the Philosophy of Science*, vol. 11 (1960–1961), pp. 212–221.

POWELL, B. "Uncharacteristic Actions," *Mind*, vol. 68 (1959), pp. 492–509.

RAAB, F. V. "Free Will and the Ambiguity of 'Could'," *The Philosophical Review*, vol. 64 (1955), pp. 60–77.

——— "The Relevance of Morals to our Denials of Responsibility" in H. Castañeda & C. Nakhnikian (eds.), *Morality and the Language of Conduct* (Detroit, 1963).

RALLS, A. "The Ascription of Personal Responsibility and Identity," *The Australasian Journal of Philosophy*, vol. 61 (1963), pp. 346–358.

RANKIN, K. W. "Doer and Doing," *Mind*, vol. 69 (1960), pp. 361–371.

——— *Choice and Chance* (Oxford, 1961).

——— "A Deterministic Windmill," *The Australasian Journal of Philosophy*, vol. 41 (1963), pp. 233–245.

RAPHAEL, D. D. *Moral Judgment* (London, 1955), Ch. X.

ROBERTS, M. *Responsibility and Practical Freedom* (Cambridge, 1965).

ROXBEE-COX, J. W. "Can I Know Beforehand What I am Going to Decide?" *The Philosophical Review*, vol. 72 (1963), pp. 88–92.

RYAN, A. "Freedom," *Philosophy*, vol. 40 (1965), pp. 93–112.

SCARROW, D. S. "On an Analysis of 'Could Have'," *Analysis*, vol. 23 (1962–1963), pp. 118–120.

SCRIVEN, M. "The Present Status of Determinism in Physics" (Symposium: Determinism in the Light of Recent Physics), *The Journal of Philosophy*, vol. 54 (1957), pp. 727–741.

SELLARS, R. W. "Guided Causality, Using Reason, and 'Free-Will'," *The Journal of Philosophy*, vol. 54 (1957), pp. 485–493.

SHUTE, C. "The Dilemma of Determinism after Seventy-five Years," *Mind*, vol. 70 (1961), 331–350.

SHWAYDER, D. S. *The Stratification of Behaviour* (London, 1965).

SILBER, J. "Human Action and the Language of Volitions," *Proceedings of the Aristotelian Society*, vol. 64 (1963–1964), pp. 199–220.

SKINNER, R. C. "Freedom of Choice," *Mind*, vol. 72 (1963), pp. 463–480.

SMART, J. J. C. "Free-Will, Praise, and Blame," *Mind*, vol. 70 (1961), pp. 291–306.

——— "Causality, and Human Behaviour," *Proceedings of the Aristotelian Society*, supplementary vol. 38 (1964), pp. 143–148.

SMITH, J. W. "Impossibility and Morals," *Mind*, vol. 70 (1961), pp. 362–375.

SPAKOVSKY von Anatol, *Freedom, Determinism, Indeterminism* (The Hague, 1963).

STENNER, A. J. "On Predicting our Future," *The Journal of Philosophy*, vol. 61 (1964), pp. 415–428.

STRAWSON, P. F. "Freedom and Resentment," *Proceedings of the British Academy*, London, vol. 48 (1962), pp. 187–211.

SWIGGART, P. "Doing and Deciding to Do," *Analysis*, vol. 23 (1962–1963), pp. 17–19.

TAYLOR, R. "I Can," *The Philosophical Review*, vol. 69 (1960), pp. 78–89 (reprinted in Morgenbesser & Walsh, *Free Will* [1962]).

——— *Metaphysics* (Englewood Cliffs, 1963), Chs. 4–5.

——— "Deliberation and Foreknowledge," *American Philosophical Quarterly*, vol. I (1964.1), pp. 73–80.

——— "Not Trying to do the Impossible," *The Australasian Journal of Philosophy*, vol. 42 (1964.2), pp. 98–100.

——— *see* Lehrer, K. and Taylor, R. (1965).

THALBERG, I. "Abilities and Ifs," *Analysis*, vol. 22 (1961–1962), pp. 121–126.

——— "Foreknowledge and Decisions in Advance," *Analysis*, vol. 24 (1963–1964), pp. 49–54.

——— "Freedom of Action and Freedom of Will," *The Journal of Philosophy*, vol. 61 (1964), pp. 405–415.

THOMAS, G. "Comments: Abilities and Physiology," *The Journal of Philosophy*, vol. 61 (1964), pp. 321–328.

VESEY, G. N. A., "Volition," *Philosophy*, vol. 36 (1961), pp. 352–365.

VIVIAN, F. *Human Freedom and Responsibility* (London, 1964).

WAISMANN, F. "The Decline and Fall of Causality" in A. Crombie (ed.), *Turning Points in Physics* (Amsterdam, 1959).

WALSH, J., *see* Morgenbesser, S. and Walsh, J. (1962).

WEISSMANN, H. A. "Freedom in Ethics—A Transcausal Concept," *Philosophy and Phenomenological Research*, vol. 19 (1958–1959), pp. 341–353.

WHITELEY, C. H. "Can," *Analysis*, vol. 23 (1962–1963), pp. 91–93.

WHITTIER, D. "Causality and the Self," *The Monist*, vol. 49 (1965), pp. 290–303.

WILD, J. *Existence and the World of Freedom* (Englewood Cliffs, 1963), Part III.

WILL, F. L. "Intention, Error, and Responsibility," *The Journal of Philosophy*, vol. 61 (1964), pp. 171–179.

WILLIAMS, C. J. F. "Logical Indeterminacy and Freewill," *Analysis*, vol. 21 (1960–1961), pp. 12–13.

WILLIAMS, G. "The Natural Causation of Human Freedom," *Philosophy and Phenomenological Research*, vol. 19 (1958–1959), pp. 529–531.

WILSON, J. "Freedom and Compulsion," *Mind*, vol. 67 (1958), pp. 60–69.

WOOTTON, B. *Social Science and Social Pathology* (London, 1959), Ch. VIII.

YOLTON, J. W. "Act and Circumstance," *The Journal of Philosophy*, vol. 59 (1962), pp. 337–350.

ZINK, S. *The Concepts of Ethics* (London, 1962), Chs. VI–VIII.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 3, July 1967

II. PERSONS AND PSYCHOLOGICAL CONCEPTS

ROBERT C. COBURN

THIS paper has two objectives: to state a view concerning the nature of persons and to argue that this view is correct. The view in question is essentially an amended version of the view set forth by P. F. Strawson in Ch. 3 of *Individuals*. It will, accordingly, be convenient to begin by stating the basic features of Strawson's view and indicating why and how this view stands in need of amendment.[1]

I

2. According to Strawson, persons are "material substances" of a certain kind. That is, they are (a) particulars, (b) particulars which endure through time maintaining their identity through change, and (c) "physical" particulars, i.e., particulars which (*inter alia*) occupy space throughout every temporal stretch of their existence and are perceivable visually, tactually, etc. What distinguishes the material substances which are persons from those which are not, moreover, is essentially that persons paradigmatically instantiate not only concepts expressible by such phrases as "is at such and such a place," "weighs so and so many pounds," and the like—concepts which, as Strawson puts it, we apply to objects of a kind "to which we would not dream of . . . ascribing states of consciousness" (p. 104); but also concepts of quite a different kind, namely, concepts which are such that (i) there are "criteria of a logically adequate kind" (p. 107) for their application which (often) can be known to be satisfied on the basis of observation of the behavior of those to whom they apply (see p. 108), and (ii) they can be self-applied "otherwise than on the basis of observation" (p. 108) and/or on a quite different (non-observational) basis (see pp. 107 and 110).[2] Concepts of this latter kind Strawson speaks of as a subclass of those which "imply the possession of consciousness on the part of that to which they are [applied]" (p. 105; see also pp. 106ff.), i.e., as a subclass of those concepts which are more familiarly referred to as "psychological" or "mental" concepts, and which include *inter alia* sensation-concepts (like *pain*, *itches*, and *nausea*), perception-concepts (like *seeing*, *hearing*, and *smelling*), action-concepts (like *intending*, *talking*, and *taking a walk*), emotion-concepts (like *anger*, *fear*, and *grief*), and epistemic-concepts (like *knowing*, *believing*, and *remembering*).[3] Finally, Strawson also maintains, if I understand him correctly, that it is logically ineluctable that things such as we are (anyhow more or less) should think of themselves in the way just indicated, i.e., as material substances which instantiate (*inter alia*) at least some psi-concepts. At any rate this is, I think, a plausible interpretation of what Strawson's talk of the "primitiveness" of the concept of a person comes to when taken in conjunction with his argumentation that we (and

---

[1] All page references in parantheses are to P. F. Strawson, *Individuals* (London, 1959). In direct quotations, I have taken the liberty of altering Strawson's spelling at the points where American English diverges from English English.

[2] Strawson, I should note, seems to say "and" and not "and/or." I have used "and/or" here to make it possible for me to refer to concepts satisfying (i) and (ii) by means of the same term I subsequently define in such a way that the concepts to which the term applies do not satisfy the second clause of (ii). Also, I shall throughout understand the notion of a concept's being self-applicable "otherwise than on the basis of observation," or as I later say "independently of observation and/or inference," as the notion of a concept which can upon occasion at least be self-applied correctly, confidently (i.e., with confidence that one's application of the concept in question is correct) and above all "with right," even though such self-applications are not, or are not generally, based upon, or justifiable by reference to, anything observed or anything inferred from observation. The expression "with right" I derive from Wittgenstein's remark in sec. 289 of the *Philosophical Investigations* (Oxford, 1953) that "to use a word without a justification does not mean to use it without right"; thus to say that one can (upon occasion at least) correctly and confidently and "independently of observation and/or inference" self-apply some concept *with right* is to say that one's confidence is perfectly appropriate or rational, even though one's application of the concept in question is not justifiable by reference to observation and/or inference. (I shall advert to this point in connection with the augumentation of para. 37 below.)

[3] Hereafter I shall refer to those psychological concepts of which both (i) and (ii) are true—assuming there are some—as *psi-concepts*. This is not a definition. A definition of the notion of a psi-concept is given in para. 19 below.

presumably things not "utterly unlike" us [p. 113]) do and must operate with some such concept.[4]

3. That this view stands in need of amendment, as I have charged above, becomes especially evident in the light of Strawson's treatment of two problems to which his view quite naturally gives rise. The first of these problems is how the fact that some psychological concepts are self-applicable on bases quite distinct from those on which one applies such concepts to others can be "reconciled" with the fact that these concepts (or anyhow some of them) are applicable to others on the basis of observable phenomena (or types of behavior, Strawson suggests [see p. 108]) which are "logically adequate" for their application.

4. Exactly why there is a problem here Strawson doesn't spell out in detail. However, in view of certain remarks he makes in this connection concerning "some of the familiar philosophical difficulties in the topic of mind" (p. 109; see also p. 110), it is not difficult to guess the kinds of considerations which lay at the back of his mind when he raised this problem of "reconciliation." Such considerations probably were the following. (1) To hold that there exist logically adequate behavioral criteria for the application of certain psychological concepts is in effect to deny the existence of a "logical gap" (p. 110) between certain sorts of behavior and the characteristics ascribed by applying the concepts in question. But to deny the existence of such a logical gap in this instance is, it would seem, tantamount to identifying the sorts of behavior in question with the characteristics ascribed on the basis of observation of these sorts of behavior. Acceptance of such "philosophical behaviorism" (p. 109) as regards the psychological concepts for which there are logically adequate behavioral criteria, however, would appear to be incompatible with maintaining that such concepts are self-applicable on non-observational bases: after all, how could one be in a position to apply to oneself independently of observation a concept which is properly analyzed in completely behavioral terms? Moreover, (2) to hold that some psychological concepts are self-applicable on non-observational bases is in effect to hold that it is at least conceivable that all or most applications of such concepts to others on the basis of observations of their behavior should be mistaken. After all, if I can say correctly, confidently, and "with right" (i.e., perfectly appropriately) on non-observational grounds that such and such a psychological concept applies to me, how could your judgment on observational grounds that it doesn't be correct? But acceptance of such "philosophical scepticism" (p. 109) as regards the application of the psychological concepts which are self-applicable on non-observational bases would appear to be incompatible with maintaining that there are logically adequate behavioral criteria for the application of such concepts.

5. Whether or not these considerations are the ones which led Strawson to think there exists a problem of "reconciliation" of the sort he considers, it is true in any case, I think, that such considerations do create (at least ostensibly) such a problem. They are, however, not the only ones which can seem to give rise to such a problem; for it can also seem to arise in view of the following, somewhat different, considerations. (3) If there exist logically adequate behavioral criteria for the application of certain psychological concepts, then it might be said that a person *A* could (in principle anyhow) always know with certainty whether any such concept applies or fails to apply to a second person *B*. But if this is so, this line of argument would continue, then *A* could (in principle anyhow) always be in a position to know with certainty that *B* is mistaken in affirming or denying that such a concept applies to himself, and *a fortiori* that it cannot be the case that there are any psychological concepts which are both self-applicable (correctly, etc.) without observation and applicable on the basis of logically adequate behavioral criteria. Again, (4) if there exist psychological concepts which are self-applicable (correctly, etc.) on "entirely adequate" bases (p. 107) which are "quite distinct" (p. 107) from those on which one applies such concepts to others, then it follows that the basis for the self-application of such a concept is some sort of "private" thing, i.e., something inaccessible to third-person observation. But if *A*'s basis for self-applying a certain psychological concept (say *C*) is thus ineluctably hidden from *B*, it is difficult to see how there could be logically adequate behavioral criteria for the application of *C* to *A*, since *A*'s basis for self-applying *C* must be relevant to the correctness of the application to

---

[4] I say "anyhow more or less" and "some such concept" because Strawson does admit the logical possibility that things not too different from us should operate with a concept not very much different from the concept of a person with which we operate. I am thinking here of his discussion of the concept of a *group* person as opposed to the (our) concept of an individual person (see pp. 112–115).

himself, and *B* could never be in a position to know by observation of *A*'s behavior alone that this "basis" is present.

6. Strawson's treatment of this first problem consists, in the main, in a mere reiteration of the claim that there are psychological concepts of the kind in question, i.e., psychological concepts the problematic character of which the problem asserts, together with an insistence that the existence of such concepts be acknowledged. Thus speaking of one particular member of the class of concepts in question—the concept of depression—Strawson writes:

> We speak of behaving in a depressed way (of depressed behavior) and we also speak of feeling depressed (of a feeling of depression). One is inclined to argue that feelings can be felt but not observed, and behavior can be observed but not felt, and that therefore there must be room here to drive in a logical wedge. But the concept of depression spans the place where one wants to drive it in. We might say: in order for there to be such a concept as that of X's depression, the depression which X has, the concept must cover both what is felt, but not observed, by X, and what may be observed, but not felt, by others than X (for all values of X). But it is perhaps better to say: X's depression *is* something, one and the same thing, which is felt, but not observed, by X, and observed, but not felt, by others than X. . . . To refuse to accept this is to refuse to accept the *structure* of the language in which we talk about depression (pp. 108–109; italics his).

However, he also produces an analogy designed to help remove the obstacles which stand in the way of acknowledging the existence of the perplexing concepts in question. "If one is playing a game of cards," he writes,

> the distinctive markings of a certain card constitute a logically adequate criterion for calling it, say, the Queen of Hearts; but, in calling it this, in the context of the game, one is ascribing to it properties over and above the possession of these markings. The predicate gets its meaning from the whole structure of the game. So [he continues] with the language in which we ascribe P-predicates [i.e., apply psychological concepts, including concepts of the type in question] (p. 110).

7. This way of dealing with the problem at hand is, however, obviously inadequate. In the first place, merely to reiterate the claim that there exist psi-concepts, as I have called them, in no way reduces the perplexities which Strawson admits are difficult to avoid concerning this claim. One cannot acknowledge the existence of concepts which one feels constrained to think impossible owing to considerations of the kind I have indicated in paras. 4 and 5 above. In the second place, the card analogy by itself is clearly insufficient to undercut all of these latter considerations. For consider: there is nothing in the case of playing-cards analogous to the first-person/third-person distinction which is essential to the definition of psi-concepts. Since this is true, there is no room in the card case, as there can seem to be in the case of psi-concepts (see (2) para. 4), for the sort of conflict in judgments between one person and another as regards the presence of the properties for which there are logically adequate criteria. Nor, again, is there room in the card case, as there is in the case of psi-concepts, for the notion that one person's basis for saying that something possesses such and such a property is different from another's and (perhaps) necessarily inaccessible to this other. In view of these facts, the card analogy obviously leaves untouched several of the arguments which give rise to the perplexities this analogy is apparently meant to dissolve.

8. Despite the inadequacy of Strawson's handling of the problem he sees and explicitly raises concerning psi-concepts, this problem can be solved. The solution I wish to advocate involves, however, (*inter alia*) two modifications—or (perhaps) one refinement and one modification—of Strawson's understanding of psi-concepts. Specifically, it involves (a) understanding the notion of a "logically adequate criterion" for the application of a psi-concept in a way which is quite different from the way Strawson's card-analogy (at least) implies he understands it, and it involves (b) abandoning the idea that psi-concepts are self-applicable on bases of a non-observational kind.

9. In stating his card-analogy Strawson, as I have noted, says: "The distinctive markings of a certain card constitute a logically adequate criterion for calling it, say, the Queen of Hearts." This suggests that when Strawson uses the expression "logically adequate criterion" (and its cognates), he has at the back of his mind the notion of logical entailment. In any case, it is quite natural to interpret the expression by reference to this notion, as comes out clearly in the fact that he is just so interpreted by at least one sympathetic reviewer.[5] The different way of understanding a

---

[5] See J. O. Urmson, "Critical Notice on Strawson's *Individuals*," *Mind*, vol. 70 (1961), especially p. 261.

logically adequate criterion for the application of a psi-concept to which I have alluded is as a phenomenon or circumstance (or set of phenomena or circumstances) which constitutes evidence that the concept in question applies in virtue of the very structure of the concept, and which hence constitutes "noninductive evidence" that the concept in question applies.[6] (This evidence furthermore may or may not be "conclusive," where evidence that p is *conclusive* if and only if it is logically impossible that a person apprised of this evidence should doubt whether p is true.)

10. It can readily be shown that understanding the notion of a logically adequate criterion in this way—together with a relinquishment of the idea that psi-concepts are self-applicable on bases of a special non-observational kind—almost suffices to make possible a satisfactory resolution of the problem under consideration. To begin with, once it is recognized that any observable phenomenon (or set of phenomena) which constitutes a logically adequate criterion for the application of some psi-concept does not *eo ipso* entail this concept's applicability to the person in question, (i) the supposition that acknowledging such criteria is tantamount to accepting "philosophical behaviorism" becomes gratuitous, and (ii) the same is true as regards the notion that acknowledging such criteria means that all first-person applications of psi-concepts are necessarily corrigible (at least in principle) on the basis of third-person criteria. Insofar as this is true, the considerations constituting arguments (1) and (3) in paras. 4 and 5 above are rendered quite harmless. Moreover, the fourth argument indicated above (in para. 5), since it turns upon the existence of non-observational bases for the self-application of psi-concepts, obviously falls apart as soon as the claim that such bases exist is abandoned. This leaves as a serious source of the first problem Strawson raises concerning psi-concepts only argument (2) (para. 4 above). This argument is, however, best treated in connection with the second problem Strawson raises concerning his account of the concept of a person.

11. This second problem which Strawson raises —consideration of which, as we shall see, necessitates still another amendment of his view—is the problem as to how psi-concepts are possible, or, as he alternatively expresses it, the problem of "what it is in the natural facts that makes it intelligible that we should have this concept" (p. 111), i.e., the concept of a material substance to which, *inter alia*, psi-concepts are applicable.

12. Given his own understanding of psi-concepts and through them of the concept of a person, it is of course not difficult to see why this second problem should present itself. Moreover, the problem is by no means dissolved, or even diminished in force, once Strawson's view has been amended in the ways indicated in paras. 8 and 9 above. Indeed, one might even feel that these amendments are of no help at all. The place of the problem they go some distance toward dissolving is, in any case, immediately filled by two other equally serious ones, each of which is a special form of this second problem Strawson raises. These are: (a) the problem as to how it is possible that there should be concepts of such a kind that there can (in principle) be evidence for their application which is paradoxically neither inductive in the broad sense (i.e., based on experience) nor deductive (i.e., such as logically to entail the proposition(s) for which it is evidence); and (b) the problem as to how it is possible that there should be concepts of such a kind that they can be self-applied correctly, confidently, and "with right," even though their application is not based upon observation or indeed upon anything at all.[7]

13. Strawson treats this second problem he raises concerning his view of persons by doing essentially two things. First, he points out that there are some psi-concepts which are such that "one feels minimal reluctance to concede that what is ascribed in these two different ways [i.e., when they are self-applied and when they are other-applied] is the same" (p. 111), owing to "the marked dominance of a fairly definite pattern of bodily movement in what they ascribe, and the marked absence of any distinctive experience" (p. 111). (Examples of psi-concepts of the kind in question are *going for a walk*, *coiling a rope*, *playing ball*, and *writing a letter*.) Keeping this fact in mind, he then suggests, can

---

[6] The expression "noninductive evidence" is Sydney Shoemaker's: see his *Self-Knowledge and Self-Identity* (Ithaca, New York, 1963), especially pp. 3 ff. The idea, however, as Shoemaker notes, derives from the later Wittgenstein. For the conception of psi-concepts developed in this paper—and defined in para. 19 below—I am also much indebted to Shoemaker's treatment of the concept of personal identity (and related concepts) in the book mentioned. The basic ideas here too, however, derive ultimately from the *Investigations*.

[7] Problem (a) is given powerful expression in A. J. Ayer, *The Concept of a Person* (London, 1963), pp. 95–96; problem (b) is expressed, in one form anyway, by G. N. A. Vesey in "Baier on Vesey on the Place of a Pain," *The Philosophical Quarterly*, vol. 15 (1958), p. 64.

"release us from the idea that the only things we can know about without observation or inference, or both, are private experiences" (p. 111) and hence from the idea that we could not possibly have "knowledge without observation and/or inference" about things which are also things which we can know about by observation and/or inference. Secondly, he tries to suggest what some of the "natural facts" are upon which the existence of the psi-concepts we employ rests by sketching some conceivable circumstances in which it would be natural to employ in connection with things importantly like us psychological concepts different in certain ways from those we in fact apply to ourselves. Thus he writes at one point:

> . . . suppose that there is a class of moving natural objects, divided into groups, each group exhibiting the same characteristic pattern of activity. Within each group there are certain differentiations of appearance accompanying differentiations of function, and in particular there is one member of each group with a distinctive appearance. Cannot one imagine different sets of observations which might lead us in the one case to think of the particular member as the spokesman of the group, as its mouthpiece; and in the other case to think of him as its mouth, to think of the group as a single *scattered* body? . . . [And] as soon as we adopt the latter way of thinking . . . we are no longer tempted to say: Perhaps the members have consciousness (pp. 114–115; italics his).

14. Unfortunately, this treatment of the problem under consideration is not much more adequate than the treatment Strawson provides of the first problem he raises. This is so for a number of reasons. To begin with, the first move which Strawson makes—that of drawing attention to certain psi-concepts like *coiling a rope* and *writing a letter* which display features quite different from those of *pain* and *anger*—simply does not bear upon the problem as to "what it is in the natural facts that make it intelligible that we should have [the concept of an individual person as he articulates it]." In mitigation of this charge of irrelevance, it might be said that the question "How are psi-concepts possible?" can of course be construed in such a way that the move in question *is* relevant, and that perhaps Strawson had some such interpretation at the back of his mind when he undertook to make it. But however this may be, the fact remains that on the interpretation of the question "How are psi-concepts possible?" which Strawson explicitly gives (see para. 11) removing a prejudice which might stand in the way of acknowledging the existence of psi-concepts goes no way toward answering the question alleged to be under consideration. Moreover, the second move Strawson makes—that of suggesting some conceivable circumstances in which it would be natural to employ psychological concepts quite different from those we continually use in talking of ourselves—is, though clearly relevant to the problem he poses, nonetheless inadequate because of its sketchiness. Not only does Strawson fail to indicate exactly what observations might lead us to think of the "spokesman of the group" as its mouth (etc.), but also, and more important, an explicit account of such observations would at best only make clear some of the facts which underlie our employment of the concept of an *individual* person as opposed to that of a *group* person: we still would have the problem as to what facts underlie our possession of the concept of a person when this concept is understood in the generic way noted in para. 2. Moreover, this inadequacy is italicized once the problem is seen to include problems (a) and (b) noted in para. 12.

15. Once again, however, despite Strawson's failure adequately to resolve this second problem he raises, the problem is by no means intractable. Indeed, to resolve it, as well as the sub-problems (a) and (b) to which it gives rise in view of the two amendments I have advocated above, it suffices (I think) simply to call to mind the general point made by Wittgenstein that the language games we play (or the concepts we employ) "rest upon" certain contingent facts concerning human beings and/or the world in the sense that if certain facts were otherwise, it would be logically impossible for us to play these language games (or employ these concepts).[8] For once this point is recalled and

---

[8] This point is made in numerous ways and places in the *Investigations;* perhaps most obviously in Wittgenstein's insistence that our ability to employ general terms at all (and so to possess the concepts they express) "rests upon" (in the sense indicated) the fact that, given a certain training, we do "go on in the same way," i.e., generally agree in our application of these terms (concepts). (See, e.g., *Philosophical Investigations,* sects. 208 ff., especially secs. 240, 241, and 242.) The "facts" upon which our concepts "rest" are ultimately, as I say, contingent ones; however, it should be noted that sometimes it is convenient to speak of concepts as "resting upon" facts the statement of which involves employment of the concepts in question and which hence are non-contingent. For brevity's sake, I employ this device immediately below (and in paras. 29 and 34) when I speak of the possibility of the concept of pain's "resting upon" a general correlation between the existence of pains and bodily and/or behavioral facts of a certain kind. (Cf. in this connection Shoemaker, *Self-Knowledge and Self-Identity, op. cit.,* pp. 229–243.)

reflected upon, the idea that there should exist concepts such that (a) there can (in principle) be "noninductive" observational evidence of their application and (b) they can be correctly, confidently, and "with right" self-applied even though their application is not based either upon observation or upon anything else, ceases to be puzzling in virtue of the possibility that the existence of certain regular (though not necessarily invariable) connections holding between phenomena of one sort and phenomena of a quite different sort should underlie (in the sense indicated) our employment of psi-concepts. Thus suppose it were the case that there existed a certain regular connection between bodily and/or behavioral facts of some kind and a "mental" phenomenon like being in pain, and suppose that this connection were a logically necessary condition of our having the concept of pain which we do have; then clearly this concept would satisfy description (a) because presence of bodily and/or behavioral facts of the kind in question would constitute noninductive observational grounds for ascribing pain to someone in view of the logical connection between our having the concept and the holding of the general correlation between the existence of pain and observational facts of the kind in question. And suppose it were the case further that there existed a certain regular connection between bodily and/or behavioral facts of the above kind and an irresistible tendency or disposition on the part of those subjected to a certain linguistic training confidently to self-apply the concept of pain, and suppose that this connection also were a logically necessary condition of our having the concept of pain which we do have; then this concept would satisfy description (b) as well as description (a) because it would necessarily be the case that a person having this concept will confidently self-apply it independently of observation, whenever he does so the concept will—at least generally—apply, and *a fortiori* he will when he self-applies it apply it "with right."

16. Once this way of solving the second problem Strawson raises is recognized, however, it immediately becomes clear why the second consideration noted above as underlying Strawson's first problem is impotent, and it also immediately becomes clear that another amendment of Strawson's general view is now required. The consideration in question, it will be recalled, turned upon the ostensible difficulty in seeing how it could possibly be the case that we might be able to self-apply some concept independently of observation without this fact's entailing scepticism as regards application of the concept to others on the basis of observation. In light of the way of dealing with Strawson's second problem presented in the preceding paragraph, however, this difficulty is clearly an illusion: for this solution makes clear that there is no logical absurdity in the idea that some concepts should be self-applicable correctly, confidently, and "with right" independently of observation even though governed by "outward" criteria of a logically adequate kind—in the sense I have given to these words above.

17. That this solution requires still a third amendment of Strawson's general view, moreover, can be seen from the following considerations. If there are psi-concepts, and if further the possibility of the existence of such concepts does depend (in the sense indicated) upon the existence of general connections of the types noted in para. 15, then it follows that such concepts are governed not only by "bodily and/or behavioral" criteria of the kind of which pain-behavior (wincing, groaning, and the like) is often taken as paradigmatic, but also by "verbal behavior" of the kind which is often emitted when a person confidently applies such a concept to himself—and *a fortiori* also by the behavioral *etc.* criteria which justify ascribing judgments or awareness of the sort such "verbal behavior" characteristically expresses. For if it is a logically necessary condition of our having and employing psi-concepts, like *pain*, that, at least generally, whenever a person confidently says (or is aware that) such a concept applies to himself, it does so, and conversely; then it follows that, for example, a person's confidently saying that he is in pain, i.e., saying that he is in pain with full confidence that what he says is true, is noninductive evidence that he is in pain.

18. Thus far I have stated what I take to be the essential features of Strawson's view concerning the nature of persons; I have stated two problems to which this view naturally gives rise and I have indicated the inadequacy of Strawson's treatment of these problems; finally I have shown how both problems can be resolved and I have noted how these approaches to the problems under consideration involve amending at three points Strawson's conception of psi-concepts and hence of persons. I want now, by way of concluding this first part of the essay, to state in summary form the view concerning persons which results when the three amendments I have noted are built into the view Strawson presents, and thereby to state succinctly

the view which in the following part I shall argue to be true.

19. In sum then: (1) persons, according to the view for which I shall argue, are material substances in the sense of this expression indicated above (see para. 2). Moreover, (2) they are material substances of a kind to which (*inter alia*) psi-concepts are applicable, where a concept C is a *psi-concept* if, and only if, it satisfies the following four conditions: (a) C's applicability to a material substance, say P, can be asserted and/or maintained by P correctly, confidently, and "with right" independently of any observations and/or inferences on P's part; (b) C's applicability to P can, in principle, be known by other material substances of the kind in question on the basis of observation and/or inference from observation; (c) some kinds of observable phenomena are such that their presence constitutes "noninductive" (and possibly but not necessarily "conclusive") evidence that C applies to P; and (d) one of these kinds of observable phenomena consists in P's confidently asserting that C applies to P. Finally, (3) persons are such that it is logically impossible that they should think of themselves in such a way that (1) and (2) fail to hold; or, more fully, they are such that it is logically ineluctable that they should apply to themselves at least some psi-concepts and also that in so doing they should apply these concepts to, and only to, things which count as material substances.

II

20. I turn now to the task of showing that the view summarized in para. 19 is true. It will be convenient to proceed by considering the three ingredients of this view *seriatim*.

21. It is, I think, obvious that material substances are included among the things to which we do in fact apply the concept *person*—to which we do in fact refer when we refer to persons, and to which we do in fact attribute (*inter alia*) hopes and thoughts, moral obligations and religious beliefs. Not all material substances to be sure; indeed not even all material substances to which expressions like "human body" correctly apply: corpses after all, though human bodies, are not persons, but *former* persons. The history of philosophy makes evident, of course, that the obvious is not beyond denial; and the history of science makes evident that the obvious is often rightly denied. However, whenever one denies the obvious one does so at the peril of laying one's arguments open to reversal; and in philosophy at least, as Moore's work has shown, this peril is a considerable one. That we apply the concept *person only* to material substances is perhaps less obvious. However, I shall argue below (paras. 41 ff.) that we cannot intelligibly apply the concept to things which do not qualify as material substances. And since (1) of para. 19 is a good deal less controversial than both (2) and (3), I shall assume that these brief remarks, together with the argumentation below to which I have alluded, suffice for its recommendation.

22. That any of the concepts we in fact and quite correctly apply to persons (and *a fortiori* to the same things to which we apply the concept *person*) are psi-concepts as defined in para. 19 is, as I say, at the very least controversial. Accordingly, argumentation is called for here of a kind which, if supplied at length for (1), might appear otiose. Since a concept is a psi-concept if and only if four conditions are satisfied, the question whether any of the concepts we employ are psi-concepts can conveniently be divided into four, each concerning one of these conditions.

23. The first question we must consider, then, is the question whether any of our concepts are such that we can upon occasion apply them to ourselves correctly, confidently, and "with right" without its being true that our basis for self-applying them is observation and/or inference. One might feel, of course, that the answer to this question is so obviously affirmative that no more need be said here than I said concerning (1). In view of the fact, however, that this answer has so recently been questioned and/or so often overlooked,[9] it will not, perhaps, be out of place here to present some considerations designed to italicize the obvious.

24. Consider the following concepts: *pain*, *belief*, *hatred*, (*intentional*) *action*, *thought*, *memory*. That each of these concepts can be correctly applied to persons is clear from such considerations as the following. In the first place, if someone were "out of the blue" to wonder or express serious doubt whether human beings ever hate other human beings, or ever believe things, or ever feel pain, etc., then the natural reaction of a philosophically untutored (or uncorrupted) person would be—at least under most circumstances—puzzlement, a failure to understand the person in question. And

---

[9] See in this connection N. Malcolm's paper "Behaviorism as a Philosophy of Psychology" in T. W. Wann (ed.), *Behaviorism and Phenomenology* (Chicago, 1964), Sect. 2.

if argumentation were presented to substantiate such a doubt or make such wonderings intelligible and seemingly legitimate, it would obviously be subject to reversal *à la* Moore, owing to the certainty of the proposition that we do upon occasion correctly attribute to persons hates, beliefs, pains, etc. Moreover, if a tribe were encountered the members of which employed certain terms in such a way that it was quite clear that the concepts these terms expressed were never correctly applied to the members of the tribe or any other human beings, then this fact would show decisively that the concepts in question were not the same as our concepts of thought, belief, pain, memory, etc. Finally, if a child were to restrict his use of expressions like "is in pain," "is angry," "believes that such and such," etc., to his/her dolls and perhaps other inanimate things, showing no awareness of the fact that they can be correctly used in connection with other human beings, then this fact would give us excellent reason to suppose that the child had not yet learned the meanings of the expressions in question.

25. But now, this line of reasoning continues, parallel considerations can be produced to underscore the equally obvious facts that (i) the concepts in question can not only be correctly applied to persons but can also be self-applied correctly, confidently, and "with right," and that (ii) they can be so self-applied upon occasion without its being true that they are so applied on an observational and/or an inferential basis. Thus if, for example, someone "out of the blue" expressed doubt whether anyone could ever, under any circumstances, correctly, confidently, and "with right" say of himself that he believed his car was parked on 59th Street, a "normal" auditor would simply be at a loss as to how to take such words. Again, if for example, the members of some exotic tribe were to employ a certain word or phrase in such a way that it *always* seemed to them to be appropriate and/or to make sense for a member of the tribe to ask another upon what observations or inferences he based his self-application of the concept this word or phrase expressed, then this fact would count decisively against our translating the word or phrase in question by such an expression as "believe it will be sunny tomorrow" or "feel pain in my left leg" or "remember having done it," etc.

26. To see that the next question to be considered—the question whether any of the concepts which are self-applicable independently of observation and inference are also concepts which can upon occasion be known to have application to persons on the basis of observation and/or inference—can be answered affirmatively too, it suffices to remind oneself (i) that we do in fact often speak of other persons as hating this or that, believing such and such, being in pain, etc., (ii) that we can often justify speaking in these ways by appeal to observable facts (and/or inference from such facts), and (iii) that often it is perfectly correct to use the word "know" in connection with assertions of the kind referred to in (i). However, should such reminders appear for one reason or another less than conclusive, one obviously could supplement them by arguments parallel to those just advanced in paras. 24 and 25.

27. Moreover, in addition to arguments of this latter kind, the following argument, suggested in part by certain remarks of Strawson's, may also be urged in support of an affirmative answer to this second question. If we never know on the basis of observation and/or inference that concepts of the kind in question are applicable to others, then it would seem to follow that we never know *in any way* that such concepts are applicable to others. For how else than by observation and/or inference could we come to know such things? But if we are never in a position to know that such concepts are applicable to others, then it would seem to follow that we simply do not understand what conditions would have to hold for us legitimately to claim to know that such concepts apply to others. After all, if nothing we have ever observed, or could properly infer from what we have observed, concerning the behavior and circumstances of others could justify us in claiming to know that, e.g., *A* was in pain or that *B* believed such and such, how could we possibly be said to understand what would justify us? But now if it is true that we simply have no conception of the occasions on which we could justifiably claim to know that such concepts apply to others, it would seem that we could not even apply such concepts to ourselves. This is so because, as Strawson puts it, "The idea of a predicate [a concept] is correlative with that of a *range* of distinguishable individuals of which the predicate [or concept] can be significantly, though not necessarily truly, affirmed [applied]" (p. 99, n. 1); and one could not be said to have the idea of such a range of individuals as regards any concept *C* if one had no conception of the circumstances under which it would be appropriate to apply *C* to any individual other than oneself. But clearly, this argument concludes, we can and do self-apply such

concepts; hence it must be false that we never know on the basis of observation and/or inference that such concepts are applicable to others.

28. The third question which requires attention if it is to be shown that at least some concepts of the kind in question are psi-concepts is the question whether any of the observable phenomena upon which we base our application of such concepts to others constitutes what I have called (following Shoemaker) noninductive evidence of their application. The following argument shows, in my opinion, that this question can also be answered "yes."

29. If the existence of certain general correlations between (a) the presence of (say) pains of a certain kind or beliefs of a certain kind, on the one hand, and (b) certain kinds of phenomena the presence of which can be determined by observation and/or inference from observational data (behavior, bodily reactions, etc.), on the other, is a logically necessary condition of the existence of such concepts as *pain* and *belief*; then clearly such (b)-phenomena will constitute noninductive evidence for the application of these concepts, which is to say (*inter alia*) that the evidential force of such phenomena will not rest upon empirical discoveries *of any kind*—assuming anyhow the (I think) obvious point that we (speakers of English, for example) neither have to discover by empirical investigation that we possess the concepts of (say) pain and belief, nor what is logically entailed by our possession of these concepts. That the protasis of the foregoing statement is true, moreover, can readily be seen by noting the absurdities yielded by its denial. Thus suppose that it were not a logically necessary condition of the existence of the concept of pain that people suffering pain generally engage in "pain-behavior" of one sort or another (groan, wince, favor the painful part perhaps ever so slightly, show unusual irritability, exhibit mild discomfort reactions, etc.). Then the concept of pain would presumably be a concept which, though (i) self-applicable independently of observation or inference, was (ii) other-applicable on the basis of "inductive" evidence, if other-applicable at all. However, if both (i) and (ii) held of the concept of pain, the following absurdities would result.

30. To begin with, it would be true that the concept of pain could never be justifiably other-applied at all, in contradiction of the conclusion established in paras. 26 and 27 above. This is so because "pain-behavior" of the kinds indicated could constitute "inductive" evidence of pain only provided either (a) it were possible to note a general correlation in one's own case between the experience of pain and the presence of such "pain-behavior" prior to being in a position to ascribe pain to others, and then legitimately generalize to the case of others, or (b) the concept of pain were like the concept of, for example, an electron, i.e., if it were an ingredient of a theory of such a kind that the occurrence of "pain-behavior" would be related to belief in the presence of pain in roughly the same way in which the occurrence of tracks on the glass surface of a Wilson cloud chamber is to belief that certain kinds of charged particles have recently moved through the chamber.[10] It can be seen upon reflection, however, that neither (a) nor (b) can be defended. The main reason that (a) fails is, as noted in para. 27 above, that we could only self-apply the concept of pain provided we already knew under what conceivable circumstances we could other-apply it. However, further difficulties with (a), should the foregoing objections be liable to attack, are quite readily forthcoming. Perhaps the most obvious of them is that the generalization in question would be too feebly based to engender rational confidence.

31. Nor is it any easier to uphold (b). For consider: if the concept of pain were the ingredient of a theory in roughly the way in which the concept of an electron is, and if accordingly "pain-behavior" were evidence of pain in roughly the way in which a track on the glass surface of a Wilson cloud chamber is evidence of the presence and movement of a sub-atomic charged particle, then it would be conceivable that we should one day have good reason for giving up the theory, and this even despite the fact that the observable phenomena to be explained by it remain the same. (One theory is, after all, sometimes better than another, not because the first alone explains some new phenomenon which the second cannot handle, but simply because it is "simpler" or is more harmoniously related to other theories which developed in connection with an entirely different domain of

---

[10] A partial and somewhat hesitant defense of (b) is presented in C. S. Chihara and J. A. Fodor, "Operationalism and Ordinary Language: A Critique of Wittgenstein," *American Philosophical Quarterly*, vol. 2 (1965), sects. VIII ff. See also in this connection P. Feyerabend, "Explanation, Reduction, and Empiricism" in H. Feigl and G. Maxwell (eds.), *Minnesota Studies in the Philosophy of Science*, Vol. III, and P. Feyerabend, "Materialism and the Mind-Body Problem," *Review of Metaphysics*, vol. 8 (1963).

phenomena.) But surely what would have to be conceivable if (b) were true is not conceivable. Surely it is not conceivable that we should one day have good reason for maintaining that people have never been in pain. For if this were so, we could now wonder (i) whether anyone has ever been in pain and (ii) whether we are now really in pain, despite our very strong inclination to think so. Just as we can now wonder whether there really are such things as mu-mesons and neutrinos, despite the fact that the only currently viable explanation of certain phenomena involves the postulation of such things. But the idea that I should now wonder either (i) or (ii) is, it seems to me, no more sensible than that I should now wonder whether 2 and 2 are really 4 or whether the prime number between 20 and 25 is identical with the color of this cup.

32. If, moreover, such concepts as the concept of pain could never be other-applied with warrant, as the immediately preceding considerations imply would be true on such suppositions as the one in question, then a number of further absurdities appear to follow. First, it would seem to follow that we could never be sure that such concepts are inapplicable to any things other than ourselves. After all, if nothing we can observe could count as evidence that such a concept applies to some other thing, how could anything we could observe show that these concepts fail to apply to such a thing? But if this is so, then we could sensibly wonder whether trees and rocks are in pain, believe things, remember things, etc. Once again, however, such wonderment seems logically impossible.[11] Second, it appears to follow that we could never be sure that anyone else understands the meanings of words like "pain," "belief," etc., since we could only be sure that someone had a grasp of such concepts provided we could know that he sincerely and confidently applied them to himself when and only when he did in fact instantiate them. But it is surely absurd to hold that we can never be sure that others understand words like "pain," "believe," etc.[12] Third, it seems to follow that no one could teach another the meanings of words like "pain," "believe," and the like. At any rate, it is difficult to see how this might be done: what procedures could a person possibly employ, given the fact that he could never be sure that his tutee was in pain, etc., and he *could* be sure that his tutee could never be sure that he or anyone else was in pain, etc.? But a view which entails that one could never teach another what pain is by pricking him with a needle is *prima facie* absurd.

33. Fourth and lastly, it would seem to follow that one could come to possess concepts like *pain*, *belief*, etc., only by giving oneself a "private ostensive definition," i.e., by turning one's attention upon, for example, a pain currently experienced and resolving thereafter to call such things "pains." But this notion of how we do and must come to possess concepts like pain is in the first place false and in the second unintelligible.[13] It is unintelligible because, and in the sense that, nothing could conceivably show that we had kept our resolution and so that we used the word "pain," e.g., in accordance with the rule indicated—or indeed in accordance with any specifiable rule. Thus appeal to the judgments of others would be impossible in view of that fact that they could never be in a position to know we are having or have ever had the experience in question; appeal to the memory of the person concerned would seem to be ruled out owing to the difficulty in this sort of case of distinguishing between (a) his remembering correctly and (b) his merely seeming to remember, and hence to the difficulty of applying with confidence in this sort of case the concept *correctly remembers;* and finally appeal to observable phenomena such as physiological occurrences which correlate with the occurrence of "pains" in the person in question would likewise be ruled out in view of the impossibility of anyone's discovering that such a correlation holds given the difficulties just noted. But if nothing whatever could unambiguously show that the person in question had kept his resolution, then it is difficult to see that "He kept it" would have

---

[11] Cf. Wittgenstein, *Investigations,* sects. 350 ff.

[12] This point about what must be the case for us to be sure that another understands the meanings of words like "pain" I have derived from Shoemaker, *Self-Knowledge and Self-Identity, op. cit.*, p. 169; the same idea is certainly suggested, however, by Wittgenstein: see *Investigations*, sect. 315.

[13] In what follows I obviously rely upon a number of the points Wittgenstein makes in the *Investigations*, especially sects. 243 ff. and p. 207. Needless perhaps to say, the argument that the concept *correctly remembers* cannot be straightforwardly applied in cases where there is no conceivable way of distinguishing between a correct memory and one that merely seems correct to the "rememberer" gains considerably more force when it is expressed as an argument against the notion that we have no choice but to employ the technique of "private ostensive definition" for a *vast number* of psychological concepts. I suspect that Wittgenstein is led to say what he says in the diary-keeper passage (*Investigations*, sect. 258) in part because he has at the back of his mind this more powerful argument.

a clear application in this sort of case; as Wittgenstein would say, "The phenomena gravitate toward another paradigm."[14] But if "He kept it" could not be said to have clear application in this sort of case, then it would be likewise unclear that "pain" was used in accordance with any rule, and hence unclear that it could be correctly called a word with a meaning.

34. A similar argument can be constructed to show that the fourth and last question can also be answered affirmatively—the question, namely, whether any concepts of the kind in question are such that a person's asserting that one of these concepts applies to himself with confidence that what he says is true constitutes noninductive evidence that what he says *is* true. For this purpose what requires to be shown, of course, is that absurdities result from the dual supposition that (a) we possess such concepts as we have been considering (*pain*, *belief*, *thought*, *anger*, etc.), but that (b) there exist *no* general correlations between the applicability of such concepts to specific persons on specific occasions and the confident assertion by these persons that the concepts in question do apply to them. That such absurdities do follow from supposing both (a) and (b), however, is not difficult to see.

35. Consider, for example, the concept of belief. If the dual supposition mentioned were true, then given the correctness of the argumentation of the last several paragraphs, it follows that any person on any and every occasion could (logically) be judged, and responsibly judged, mistaken in asserting that he believes such and such. For it could be true—i.e., it is conceivable—that whenever a person confidently asserts of himself that he believes such and such, observable phenomena are present which constitute strong noninductive and/or, of course, inductive evidence that he fails to believe what he claims to believe. But surely it is absurd to think that any and every occasion on which a person confidently says he believes something is also an occasion on which he could (logically) be responsibly judged to be mistaken about his beliefs. Again, given the truth of the suppositions indicated, it would follow that a person could always wonder whether, e.g., he was in pain, or believed such and such or felt anger at so and so. For if, as the truth of these suppositions would entail, there is only a contingent connection between the fact that the concept of pain or belief or anger applies to *A* say and the fact that *A* has confidently asserted this to be so, then it would seem that *A* could always legitimately wonder whether, despite his feelings of confidence, the connection in question has broken down on the present occasion. But, again, it is absurd to think that any person could always legitimately wonder about such things. Finally, it follows, I think, from the truth of the suppositions indicated that a person, *A* say, could always justify an assertion to the effect that he, *A*, was just thinking about the weather or believed his car to be somewhere on Woodlawn Avenue, by appealing to the same kinds of observable phenomena to which others typically appeal when justifying their claims concerning *A*'s thoughts or beliefs. For consider: if the connection between the fact that, say, the concept of believing that such and such applies to *A* and the fact that *A* has confidently asserted this to be so is a contingent one, then presumably *A*'s knowledge that such a connection holds will—anyhow ultimately—be based on experience, experience of a general correlation between the presence in him (and others) of beliefs of some sort and a disposition on his part (or the part of others) to hold and assert with confidence that he believes such and such. But, once again, it is absurd to think that a person can always sensibly justify ascribing to himself a certain belief by saying, for example, (and *inter alia*) that he is strongly inclined to think and say that he has the belief in question.

36. Given the correctness of the first point made in para. 21 and the arguments presented in paras. 23–35, it follows that persons include material substances in the sense explained in para. 2, and that these are of a kind to which (*inter alia*) psi-concepts are applicable. I want next to argue that the notions that we—persons—might cease (or have failed) to operate with at least some psi-concepts, that we might cease (or have failed) to apply them to material substances, and that we could intelligibly apply them to things other than material substances, are logically absurd. I shall begin by arguing that it is logically ineluctable that we (or persons) should employ at least some psi-concepts.

37. Suppose we employed no psi-concepts whatever. Then it would follow, I believe, that we had no concepts which we could self-apply correctly, confidently, and "with right" independently of observation and/or inference. This is so because any concepts of the latter kind are necessarily psi-concepts, as the following considerations make evident. Suppose a concept, say *C*, is self-applicable

---

[14] Wittgenstein, *Investigations*, sect. 385.

(correctly, etc.) independently of observation and/or inference. Then it must also be other-applicable on the basis of observation and/or inference because if it were not, it would follow for reasons of a kind given in para. 27 above that it could not even be self-applied. But now if *C* is other-applicable on the basis of observation and/or inference, then it must also be the case that there are, in my sense of the expression, "logically adequate criteria" for its application to others, and that a person's confidently asserting it to be self-applicable independently of observation and/or inference is among these criteria. For if there were no such criteria, it is difficult to see how *C* could be other-applicable at all (compare paras. 29–31 above); and if a person's confidently self-applying *C* independently of observation and/or inference were not such a criterion, then a person could (logically) always rationally be judged wrong in so self-applying *C* despite correctly understanding the expressions which stand for *C*—which is, I think, absurd.

38. The force of this last argument turns, of course, as I have hinted in note 2 above, upon the notion that the concepts here in question are by definition not only self-applicable correctly and confidently independently of observation and/or inference, but also self-applicable "*with right*" independently of observation and/or inference. That this is so becomes clear if we construct in imagination some concepts which lack this feature. Thus suppose concepts of the form *weighs x lbs.* were such that people were occasionally able correctly and confidently to self-apply them independently of observation and/or inference. Then if a person were to state his weight on some occasion correctly and confidently and without making any relevant observations and/or inferences, we could say that his confidence was rational or appropriate, i.e., that he made his judgment "with right," only provided we knew that he had learned from experience that he generally judges correctly in such cases. But this means that he could generally judge incorrectly without thereby evidencing his lack of understanding of phrases of the form "weighs x lbs." In the case of concepts which are by definition such that a person can generally correctly, confidently, and "with right" self-apply them independently of observation and/or inference, however, a person couldn't be shown (or learn) that his confidence in self-applying them was inappropriate, since the appropriateness of his confidence is, as it were, built into the structure of such concepts. Hence if a person were to judge wrongly on every occasion when he self-applied such a concept, it would follow that he simply did not possess the concept in question. For it is absurd to suppose that a person's confidence in judging as he does should be entirely appropriate, entirely rational, and that he should nevertheless judge wrongly about the application of the concept to himself on every occasion when he self-applies it.

39. But now, this argument continues, if we had no concepts which we could self-apply correctly, confidently, and "with right" independently of observation and/or inference, it would follow that we could have no "empirical concepts" whatever, i.e., concepts which we apply on the basis of observation and/or inference from observation. For consider: to apply with justification an empirical concept, it is necessary (in general) not only to observe some phenomenon, but to be aware that one has observed this phenomenon. For if one were not aware that one had observed the phenomenon in question, one obviously would not be in a position to justify one's application of the concept in question on the basis of the observations one has made. But to be aware that one has observed some phenomenon, it is necessary that one be able to (i.e., be in a position then justifiably to) self-apply the concept of observing *x*. Now suppose that we could not (correctly, etc.) self-apply this concept independently of observation and/or inference. Then, obviously, we could only determine that we were observing some object *x* by finding out by means of observation and/or inference from observation—e.g., by observing that our eyes are open, that they are directed toward *x*, that nothing stands between our eyes and *x* to obstruct the line of vision, etc. But if this were so, then it follows, I believe, that the warranted application of any empirical concept would require an infinite series of observations, since we could only be sure that we had, e.g., observed that our eyes were open by making further observations, and so on *ad infinitum*. Since this latter notion is absurd, it follows that either we have and can have no empirical concepts whatever or that some of the concepts we employ are self-applicable (correctly, etc.) independently of observation and/or inference, and hence are psi-concepts. But surely, this argument concludes, it is a necessary condition of our being persons that we can possess empirical knowledge, i.e., that we can justifiably apply empirical concepts. If this is so, however, then in view of the foregoing it would seem to be a

necessary condition of our being persons that we employ at least some psi-concepts. In short, people—such as we are—necessarily employ some psi-concepts.

40. Several other arguments to this conclusion which are of the same general form can readily be constructed. Thus suppose again that we employed no psi-concepts whatever. Then, for the reasons noted, we would have no concepts which were self-applicable (correctly, etc.) independently of observation and/or inference; *a fortiori* we could not self-apply (correctly, etc.) independently of observation and/or inference concepts like *remembers x*, and *believes that p*. But this in turn entails that a person could only be sure, for example, that he remembered seeing the game on Tuesday or that he believed his cat was sick by observing and/or inferring from his observations that such and such——e.g., that the game took place on Tuesday (etc.) and that he was disposed to say things like "My cat is sick" (etc.). However, the notion that we should only be able to determine that we remembered such and such or believed such and such in this way is absurd. For if we could only be sure that we remembered that such and such on observational grounds, then we could never be sure at all; and this because (*inter alia*) it would be necessary to keep the necessary observations and/or inferences in mind (i.e., remember them) for some time after they were made in order to use them for the purpose of deciding whether we remembered that such and such—a fact which would also lead to a vicious infinite regress. Again, if we could only be sure that we believed that such and such on observational grounds, it would also be true that we could never be sure at all; for any relevant observations or inferences from observations we made would be observations or inferences which we could be sure we *believed* we had made only on the basis of further observations and/or inferences, and so on *ad infinitum*. But now surely none of the members of a reasonably large group of things could satisfy the concept of a person if none of them could be aware that they remembered past events; nor could the members of any such group individually satisfy this concept if they were all incapable of believing anything and being aware that this was so. If this is the case, however, then it follows again that persons necessarily possess at least some psi-concepts.

41. It remains now but to show the absurdity of the notion that we (or persons) should apply the psi-concepts we (or they) necessarily have to things significantly different from the things to which we (they) do in fact characteristically apply them—viz., material substances of a familiar kind. The following argument, which I shall present in two stages, hopefully suffices for this purpose—given of course a lenient interpretation of the words "significantly different."

42. The first stage of the argument in question consists of a consideration designed to show that psi-concepts must apply to enduring particulars of some kind. (I simply assume here the absurdity of the idea that persons should apply psi-concepts to non-particulars of some kind.) Suppose we never applied our psi-concepts to enduring particulars of some kind. Then, I think, it would be impossible for us to know that anyone else understands the psi-concepts we employ. For in order to know that someone else understands some psi-concept, say *C*, we must be able to find out that he generally self-applies *C* when, and only when, we are able on the basis of observation and/or inference to apply it to him. But this means it must be the case that there are a number of occasions on which he self-applies *C*; which in turn means that he must exist on these occasions and be the same individual on each of them. The latter condition is one which can be satisfied, however, only by an enduring particular. Accordingly, if we are to be in a position to know that others employ the psi-concepts we do—as we must in order to employ such concepts ourselves (see para. 27 above)—then, if the above reasoning is correct, it cannot be the case that we always self-applied our psi-concepts to things other than enduring particulars. However, the idea that we might *sometimes* apply psi-concepts to particulars which do not endure through time is equally infelicitous. For the notion that a non-enduring particular of some kind might have a concept like *belief*, say, be capable of self-applying it independently of observation and/or inference, etc., lacks clear sense. For given that such a thing could not exist on a number of occasions and be the same individual on each, there would seem to be no way in which it could conceivably be established that the thing in question did possess such a concept, was capable of self-applying it independently of observation and/or inference, etc. *A fortiori*, the individuals to which we apply our psi-concepts must include enduring particulars and only enduring particulars of some kind.

43. The first stage of our argument, if successful, rules out (*inter alia*) the possibility that we should apply our psi-concepts (a) to momentary particu-

lars of either a physical or a non-physical kind (e.g., momentary temporal slices of material objects considered as four-dimensional entities), or (b) to temporally extended but non-enduring[15] events or processes of either a physical or a non-physical kind (e.g., brain processes). In the next and second stage, I shall present some considerations which are designed to rule out the possibility that we should apply our psi-concepts to enduring particulars of a non-physical, or "spiritual," kind.

44. To understand what it is for something to be an enduring particular of some kind requires that one know at least roughly under what circumstances a thing of the kind in question can correctly be said to retain its identity through time. Alternatively, one must know at least roughly what the criteria of "particular-identity" for such things are. In the case of particulars of the kind in question, however—viz., enduring particulars which are non-physical—reflection reveals that there neither are, nor could be, any criteria of the required sort. Thus spatio-temporal continuity will not serve as a criterion of identity for such things, since such things, being non-physical, do not exist in space. Neither will "psychological similarity" or "memory." For there is nothing which could rule out the logical possibility of two distinct particulars of the kind in question being indistinguishable as regards any and every "psychological" feature (knowledge, temperament, etc.). Hence one could never justify saying that two temporally separated, independently identified particulars of this type are numerically the same particular because they share such and such psychological features. And the fact that a particular of the kind in question which was identified as existing at t_2 possessed "ostensible memories" of the states of a similar particular identified as existing at t_1 wouldn't suffice to justify such an identity-claim either. For nothing in logic rules out the possibility that there should exist at t_2 two (or any number of) such particulars which are indistinguishable as regards their "ostensible memories"; and the notion that one of these sets of "memories" might alone be veridical is vacuous inasmuch as it is obviously impossible that anyone should have the slightest basis for thinking one set rather than another to be the veridical set. But now if neither spatio-temporal continuity, nor psychological similarity, nor memory could serve as a criterion by appeal to which questions about the identity through time of particulars of the kind in question might be answered, it is difficult to see what could. If this is so, however, it is also difficult to see that the notion of an enduring particular of a non-physical kind makes clear sense.

45. If the above argumentation is sound, it follows (a) that we must apply our psi-concepts to enduring particulars of some kind and (b) that such enduring particulars could not fail to be physical in character. To have shown (a) and (b), of course, is not quite to have shown absurd the idea that we should apply the psi-concepts we necessarily have to things other than the material substances to which we do in fact characteristically apply them, since brains, for example, are, after all, enduring particulars of a physical kind, too, and so also would be the "scattered" bodies of the "group persons" Strawson envisages (pp. 112 ff.). However, (a) and (b) together, if true, do limit the possible subjects of our psi-concepts to things at least significantly like the things to which we do in fact, characteristically apply such concepts. And this result is all that our present aims require.[16]

The University of Chicago

Received April 18, 1966

---

[15] An "enduring" entity, in the sense involved here, is an entity which, at any time of its "existence," exists as a whole. Thus events do not endure because at any particular moment during which an event is taking place, only a part or phase of the event is present.

[16] I am grateful to Professor William Alston for criticisms which have enabled me to improve the paper at several points.

American Philosophical Quarterly
Volume 4, Number 3, July 1967

III. ON SEEING BODILY MOVEMENTS AS ACTIONS

VIRGIL C. ALDRICH

SPEAKING of the "observed movements of others" P. F. Strawson says that "we *see* such movements as *actions*" and that "we *interpret* them in terms of intention . . . we *see* others as" ascribing to themselves "not on the basis of observations" what "we ascribe to them on this basis." Of the observed movements of others Strawson also says that "we understand them, we *interpret* them, only by seeing them as elements in . . . plans or schemes of action. . . ." All this amounts to saying that "we *see* each other as persons," and that this is "a lot of things, but not a lot of separate and unconnected things."[1] I have italicized "see" and "interpret" in the above quotations to set the stage for my main question here: Where does the seeing end and the interpreting begin? Answering this question, or finding out what is wrong with it, will throw light on the question of "logically adequate criteria," a notion that has exposed Strawson to considerable fire.

As he puts it, it is not clear whether Strawson means to say that, strictly speaking, what a man is *doing* (his "action") is perceptible. From his statements, one might gather that it is only movements that are perceived (seen, observed), and that, in certain favored cases where the "logically adequate criteria" come into play, the (intentional) action—what the man is doing—is grasped by "interpretation" infallibly; that is, it being there senseless to suppose that the interpretation is mistaken. Construing the remarks this way, the contention is *not* that "the decision rests with perception" (Aristotle). Reports of actions, howsoever elementary, would not be judgments of perception. The decision, or the case for it, rests, rather, on an appeal to criteria that logically certify it, as an interpretation of movements that alone are observed or "seen"; not on the *perception* of the *person* in action. In this vein, we speak with the vulgar when we say we *saw* the student cheating, but are saving our philosophical face by thinking with the learned that we saw only certain movements, aptly *interpreting* these as expressing the action of cheating. Though we may *know* that this is going on in the logically certified cases, we still are not, even there, *observing* someone cheating. What the criterion establishes beyond question is the *existence*, not the perception, of the cheater. Carrying this sort of strictness all the way, we must also refrain from thinking that we ever *see* persons (people?). That it is a *student* caught in the act, a person, is also an interpretation of a figure in motion, this being all that we truly observe.

In these passages Strawson stays consistently with the notion of "seeing as." While he nowhere says that we simply see persons or their actions, he everywhere says that something is "seen as" a person or an action. Now, if one turns the gun of Wittgenstein's account of seeing-as on these assertions, the result seems at first distressing for Strawson. A condition of seeing x as y is that x be not y. That is, one does not see a knife as a knife, since it *is* a knife. One sees, or fails to see, *that* it is a knife. One may see a drawing as a rabbit, or a cloud as a camel, since in such cases you are not actually confronted by what you see the thing as. But Strawson, when he says we see something as an action or a person, clearly does not mean that we are not confronted by that action or that person, in the experience of them as "aspects" of what does confront us. He means to say that, in such cases, both the figures-in-motion that we simply "see" *and* what these are "seen as," are actually present there before us, provided the criteria in these cases are logically adequate. So this is a different sort of seeing-as from the sort that Wittgenstein featured.

But there are three things in Wittgenstein's model of seeing-as that Strawson can appeal to.

(1) Wittgenstein spoke of "aspect-blindness," with the presumption that anyone who is not afflicted with such blindness "sees" *the aspect*, in some sense of "see." In *that* sense, Strawson might

---

[1] "Persons," anthologized in *The Philosophy of Mind*, V. Chappell (ed.) (Englewood Cliffs, 1962), p. 143.

maintain, we see the action or the person. "But this isn't *seeing!*—But this *is* seeing!"[2] Both remarks, said Wittgenstein, can be given a conceptual justification.[3] So, one might argue that both the "bodily movement" and the action—the person in action—it is seen as, are "seen," though in different senses of the term. (2) As a matter of fact, there does seem to be this sort of difference or duality in modes of perception in the person-action case. We clearly do not "see" the action or the person quite as we see the "bodily movement" (Strawson's term). We see the latter as the former, in a manner that permits us, nevertheless, to suppose that both are seen. And since the movement is not *per se* the action, the concept of "seeing as" is appropriate here. (3) The relation of the aspect—what something is seen as—to that something in such cases as, say, the duck-rabbit picture is remarkably like the relation of the intentional action, or what the person is doing, to the pattern of movements that manifests it. The picture comes to life in an image, the movements in an action. Or, an image animates the picture; an intention animates the bodily movement. Thus the language of mind serves in the description of both cases. Yet, granting these three points of resemblance, there remains the glaring difference noted above. Though the body in motion is not identical with the person in action, as the drawing is neither a duck nor a rabbit, yet in the former case you are nonetheless in the presence of a person doing something, as you certainly are *not* in the presence of a duck or a rabbit in the latter.

It is such a consideration that opens the tent of the seeing-as position to the camel of scepticism of other minds or persons. The sceptical philosopher will ask: Might not the figure-in-motion that we see when we look at the putative student cheating be functioning like a three-dimensional "picture" or mobile statue? It would then be seen as a person doing wrong, without any person being there at all. We would only be supposing, from our own case, that a person animates the configuration of seen motions. And it is logically possible that this sort of thing is going on all the time, everywhere. A softer or more moderate form of such scepticism would be: the persons, the people around us, granting their existence, could be only pretending all the time, making certain intentions or feelings apparent in the pattern of their bodily movements. The feeling, the intention, can be expressively portrayed in an observable configuration of movements—the pretense—that is not backed by a person who actually has it, or is backed by one who is not actually in that state of consciousness. The actor exploits this fact, though without pretense in this stage-setting; and in other works of art we "see" feelings and intentions that are patently manifest or "embodied" without our being confronted by anything that actually "has" them. It would seem then that whatever it is that we do see, including what this is "seen as," one may, for all he *knows*, never have encountered another person in the supposed state of consciousness, since the decision about such things does *not* rest with perception, according to this argument. Of course, we are left anyway with criteria for mental predications, but none of these turns out to be "logically adequate" for any *particular* case after all, perhaps securing us only against the possibility of being mistaken wholesale or always. The sceptic is willing enough to grant this, for his point was from the beginning that in no particular case can one be sure of perceiving the action or even another person.[4] One who knows only that at least one other person must be (tenseless) around whom he has encountered somewhere, sometime, knows virtually nothing. And that exactly describes our condition, according to our sceptic, since perception never *shows* us any other person. The conclusion from these considerations is that perception ends before we get to other persons, and conception or interpretation of what we see takes us from there on to being sociable with them—if there are any others.

The non-sceptics have had (at least) two sorts of reactions to arguments like this. One of these is to contend that perception in no matter what form must not be made too much of, in the executive affair of one's transactions with realities. Perception thus tends to be replaced by "attitude," "action," "practice" as the fundamentals. We do not need, as philosophers, to rest the case of "knowledge" of persons and what they do on strictly perceptual grounds. The ground of perception is not solid—or "primitive"—enough for that. Rather, it is "action" that is the matrix, the bedrock. I think first of Hampshire and then of some pragmatists in this connection. To emphasize this is to by-pass questions about what is perceptible, thanks to the conviction that, in the matter of persons and human action, the decision rests neither with perception nor with thought (interpretation). Perception ends far short of the mark, in this area; and formal conception does not take us much farther. We connect with others in action.

---

[2] *Philosophical Investigations* (New York, 1953), p. 230.

[3] *Ibid.*

[4] See A. J. Ayer, *The Concept of a Person* (New York, 1963), p. 100.

My impression is that such an emphasis is made by philosophers who had too small a notion of perception to begin with, or just a wrong one. Moreover, an appeal to perception, properly understood, is the best way to "objectify" both experience and theory. So I prefer the other reaction of the two mentioned above. This is the attempt to show the liability of the old notion that perception ends where conception begins. They are indeed distinct and remain so, but they may mature together. We need a theory of perception that does not make *seeing* something "according to an *interpretation*" look like a "queer fact,"[5] while we restrict this notion sufficiently to exclude cases of imagining or believing or inferring something in the perceptual situation.[6] The adequate theory will also show that, given the context, there is relevant to it a mode of perception that is "primitive," and that, in context, some things will be perceptually too obvious for words, a point I have developed elsewhere.[7] By this I mean that it is *logically impossible* to make sense of a verbal report of what, in the situation, is too obvious for words. Anyway, all this is my way of saying that the decision rests with perception. Perhaps it also expresses the point of Merleau-Ponty's remark that all consciousness, in its fundamental intention, is perceptual—with the normative suggestion that this should be kept in mind by the theorist.

What Strawson says about experience of, and talk about, persons lacks the underpinning of an adequate theory of perception of persons by persons. What he does say about seeing and what is seen in this connection puts a strain on the other more sensible things he says about the logic of talk about persons and their experiences.

To help the fly out of the bottle, let us begin by considering various actions, from the simpler to the more complex, asking in which cases the person engaged in the action can be said to be "seen" doing it. But, first, a passing comment on the concept of "people" in relation to "persons," since there is a natural inclination to use them interchangeably. Is the expression "people who are not persons" self-contradictory? I think it is, but we do have a use for "peopled" and no corresponding use for "personed." For example, a place may be said to be "peopled" by strange creatures that are not persons. The "people on Mars" might be such creatures. So there may be such a thing as discovering that some people, in this sense, are not persons. But, these two concepts being what they are in our present use, one does not commit a bad mistake by using them as synonyms. They are so treated in an excellent recent essay on "The Philosophical Concept of a Human Body"[8] and I propose to follow suit. "Seeing people," then, is to be equivalent to "seeing persons."

Now as for the actions which are such that the agent can be "seen" doing them, let us consider first the difference between "walking" *simpliciter* and "taking a walk." If it is the former, the decision about its going on or not seems to rest with perception, and the observer can hardly be mistaken. But, it may be contended that such perceptual discernibility comes at a price. One can be quite sure only where walking approximates being nothing but a set of bodily movements, *not* "seen as an action." Thus walking loses the name of action, and we were concerned with the question of perceptible *actions*. "Taking a walk" seems to qualify here as an action by contrast with simply walking because, you may say, only people can do this and they can pretend to be "taking a walk" while actually doing something else, like spying. (They can't pretend to walk.) So, people cannot be simply "seen" doing this action that could be simulated. Thus where are we? We seem to be left either with perceptible movements or imperceptible actions, after all.

To escape this predicament, we need to distinguish the following: (1) going through the motions of walking, (2) walking, (3) taking a walk, (4) pretending to be taking a walk. (1) is distinguished from (2) to preserve "walking" as the concept of an action. But how, you will ask, can one go through the motions of walking without walking, and *vice versa?* Just what is this distinction, if any? That they are distinct concepts is driven home by the philosophical reminder of the fact that

[5] Wittgenstein, *op. cit.*, p. 200.

[6] One does not strictly "see" a right triangle, based on its hypotenuse, as having toppled over from a shorter side; one "imagines" this. Wittgenstein, *op cit.*, p. 207.

[7] "Too Obvious for Words," *Mind*, vol. 72 (1963), pp. 348–356.

[8] D. C. Long, *The Philosophical Review*, vol. 73 (1964), pp. 321–337.

walking is something we at first had to try to learn. A little child, or an adult after an accident, does not simply stand up and walk. Walking is an achievement. *One does not learn to walk by walking.* One learns by "going through the motions," in the phase of learning and trying to walk, usually with some assistance and guidance. (Aristotle might have made more of this sort of distinction: one learns to perform the action of generous giving not by giving generously, but at first by going through the motions, for a period of behavior that is visibly arduous; truly generous giving is an achievement that comes later. But Aristotle did say that "movement," contrasted with "activity," is "incomplete," and this is quite in line with the distinction I am making between "going through the motions" of an action and the real performance.)

An important consequence of *this* concept of "going through the motions" of A (the action), where A is an elementary action like walking, is that it certainly is not such motions that we "see as" A (Strawson). On the contrary, seeing someone going through the motions of A, in this sense, is to see that he is *not* doing A; rather, that he is trying and (perhaps) learning to do A. Moreover, and still more important, when we see somebody actually engaged in or doing A, there are no *such* motions that we are seeing, precisely because the agent is no longer going through them. Thus, if Strawson is nevertheless right about there being "bodily movements" that we do see "as the action," they must be distinguished from the motions the agent goes through *before* he is able properly to perform the action. The significant thing to notice here, at this point of our inquiry, is that ordinarily when we see someone, say, actually *walking*—not just going through the motions—what he is *doing* (walking, the action) seems so patent or salient for perception that the whole notion of "seeing bodily movements as the action" becomes inept. We see the *person* walking; we do not see his *bodily movements* as the action of walking. In one sense, we seem not to see "movements" at all, because what we are seeing is the person in action. To see a person as *doing* something (acting) is not to see him moving in this or that way. And, just as, in some circumstances, we tell by looking whether someone is in the stage of going through the motions of A, so we see whether someone is, instead, doing A. At the rudimentary level of actions like walking, there seems to be little use for the notion of "seeing as," which applies rather to more sophisticated actions like "taking a walk," where the agent, *seen walking*, may be "seen as taking a walk" or even as spying.

What qualifies simple walking as an action is, so far, its being something one first tries and learns to do, and may thereafter be voluntary. As for its being "intentional," the adult crippled by an accident "intends to walk" while he tries to do it. (Such intention is not appreciable in a child's learning.) When the adult can walk again, his walking ceases to be intentional. He just walks, intending to accomplish *something else* by walking, this being a more complex action that *is* intentional, the intention being served by the walking. Thus the going-through-the-motions is consummated in the total action of walking only when the agent is doing something else by walking—intends to get somewhere or at least accomplish *something* thereby. (See next paragraph on "locution" and "illocution.") The walking then is "in use," and is the "complete" action of walking only in such a connection, where the going through the motions are *aufgehoben* in the performance. It is in the perceptual awareness of the complete, intentional action, absorbing into itself the rudimentary action of walking, that we tend to overlook bodily motions or movements as such. We see the *walking* as ingredient in, say, taking a walk, or in spying. We do not see the *bodily movements* as the complete action. And the walking itself, as an (elementary) action, is such that to see it is already to be overlooking going-through-the-motions of walking. Thus is there a sort of hierarchy of metamorphoses, i.e., of motions (trying, learning) into the rudimentary action, and this into the complete intentional activity in which the simple action is ingredient. In the social situation, one ordinarily sees "at first" the rudimentary actions, though he may go on also to see either the going through the motions or the complete activity, under the right inducements.

What I am trying to convey is quite like what has been said (Austin) about the total linguistic act of "saying something." There we distinguish the going-through-the-motions as the "locution," the mere forming of the right words in the right order, the first stage of learning a language. (This, for Austin, is the "phatic" component of the locutionary act.) Then comes the "illocution," which is putting the sentence to the uses it was devised for, such as statement-making or question-asking—as walking has the common usefulness of getting the agent at least somewhere though not always anywhere in particular as in "taking a walk." Finally, there is the "perlocution," the accomplishing of the purpose or the "point" of making the illocution such as convincing someone—as in getting somewhere (walking) one achieves a certain end, like moon-viewing in taking a walk, or acquiring information in walking by a military installation (spying), or thinking as one paces the floor back and forth. And as a speaker fails really to be saying anything (performs no illocutionary act) in circumstances that provide no occasion for its use, nothing to be accomplished by it (perlocutionary act), so one does not really walk with respect simply to going through the motions; that is, if it is not ingredient in a complete action.

This comparison suggests two other sorts of situations, besides the learning-how-to-do sort, in which also just the motions of doing something are gone through. An adult walker may deliberately "illustrate" walking by putting on a show of it for the learning child, or to show a foreigner what the English word "walking" means. In both cases he could be said to be only going through the motions since, strictly speaking, he would not be walking. He would not be "using" the action of walking, but only "mentioning" the action by the demonstration he puts on. A person's going through the motions, in this sense *alone*, means presenting "bodily movements" (Strawson) that are to be "seen as" walking—the function of bodily parts in walking. But this (relatively rare) case is to be distinguished from the ordinary one of seeing someone walking, say, to his office. The latter is not to see certain bodily movements as walking (the action), but to see a pedestrian going somewhere, with no particular attention paid either by him or the observer to his bodily movements, because here he is really *doing* something. He is not, as one who walks, going through *any* "motions." (A queer style of walking, such as with mincing steps, may draw some attention to the "bodily movements.")

But going through the motions in this sense must be distinguished from yet another, the sense in which a man-shaped automaton may be said to go through the motions of walking. This is neither what the learner does nor the adult human being who "mentions" walking by putting on the show, without "really" walking. It does not make sense to say of the robot that it either tries (does not try) or learns (fails to learn) to walk. So the concept of walking collapses in such applications, such that the term should appear in scare-quotes if introduced nevertheless. But the important thing to notice here is that, at last, we have a *bona fide* case of a set of motions of a "physical object"—not even "bodily movements"—that are "seen as" the action of something walking. This is the effect that science fiction aims at. And, as we have noticed, this is the kind of seeing-as that Wittgenstein draws attention to in his description of the aspect-experience one has while looking at figures. In such cases, we properly say we see x as y precisely because there is the presupposition that the y that x is seen as does not really confront us. To see a small child going through the motions of walking is to see something trying or learning to walk, not to see the pattern of motions as something walking. But to see the automaton moving thus is not to see it as learning to do anything, and it *is* to see it "as walking," which entails that it is *not* walking. There is quite a difference.

This turn of the argument prompts me next to concentrate on the other concept of seeing-as that does have a use in talk about human beings in action. And this requires me to sharpen the distinction mentioned above between rudimentary actions such as walking on the one hand (2), and, on the other, the "complete," intentional actions such as taking a walk (3) and spying (4). Out of these considerations a new concept of "practical aspects" and "aspection" will emerge, and so we shall reach a conclusion about what it means to say, for example, that somebody was "seen" taking an evening stroll or "seen" spying.

The necessary distinctions can be made with the help of the concept of pretending. One cannot pretend to walk, though he can pretend to be "just taking a walk" (while he is spying). And he can pretend to be going through the motions in the various senses listed above, of trying to walk, illustrating it, etc. One who can walk can pretend to be unable to, by deliberately going through the motions of trying to walk and falling short of that mark—pretending to try to walk—as a coward might on the battle field. Or, one might pretend to be only illustrating what walking is, talking about it to draw attention to the bodily movements, as he *actually* walks, behind this veil of simulation, to the mantle piece to get the pistol there and shoot his way out. Finally, one might, given sufficient disguise and the right circumstances, deceive somebody into thinking he is a robot by going through the mechanical motions of walking like one—pretending to be an automaton. But, granting all this, one cannot pretend to walk. He walks or he doesn't walk, patently.

Thus, the case is different for the "complete" actions of which the action of walking is a component. The latter I have referred to as rudimentary or elementary. I now introduce the term "atomic action" for such constituent actions, and shall call the complete ones of which they are proper parts "molecular actions." An atomic action has no actions as parts of it. But it does have distinguishable parts. The atomic action of walking, for example, may be analyzed into bodily movements, no one of which is itself an action; it is not anything the agent *does* while walking. He can be said to be performing (doing) the bodily movements of walking only if he is "mentioning" it in a demonstration of walking, or to show that he can walk, etc. But this is not full-blown walking—walking for something, possibly just health or meditation—and it tends to disintegrate that action into components the conscious performance of which may spoil it. In short, the bodily movements and the walking are not things that you strictly *do* simultaneously, so the movements are not component *actions* of the action of walking.

But if the action is molecular, like taking an evening stroll, you certainly must walk to do this. And it is something you *do* (an action) in the *doing* of the molecular action. You do both at once, though the atomic action in this frame is usually absorbed into the molecular. That they are distinct actions, however, follows from the consideration that while you may be seen walking with no mistake about this—no room for pretense here—you may be pretending to take an evening stroll, as you would be were you actually spying on the nearby military installation. Then you would *not* be taking an evening stroll, despite appearances. You would only be going through the motions of doing this, as a camouflage concealing what you were really doing. The description of this as "taking an evening stroll in order to spy" is incorrect. So would "taking a walk" for that purpose. Rather, you just go through the motions of these, in the pretense. But you *must walk*, to pretend you are taking a walk or an evening stroll. (Spying by walking is another sort of molecular action, the analysis of which reveals a "perlocutionary" element; taking a walk involves only the "locutionary" and "illocutionary.") This is what makes "seeing" the person doing molecular actions problematic. The possibility of one's only appearing to do them brings "interpretation" of what is going on into the picture, in a way that has tempted some philosophers to suppose that it replaces perception of the action. One "sees" only the atomic actions—or, worse, only the bodily movements—so they are tempted to say, and do say. So they believe that in cases of the molecular actions, the decision rests not with perception but with inference or conception (interpretation), for the observer of the action in question.

Before answering this in favor of perception, let us remind ourselves of how natural and right the notion of "seeing" what someone is doing is in *some* such cases, as against the notion of only "interpreting what is seen" or "inferring the action from what is seen" or "only thinking you saw" in *all* such cases. Surely I *sometimes* see you taking an evening stroll, as, say, when I, an old friend, participate in the action with you after myself proposing it. And, in other circumstances, I might catch you in the act of spying, which would certainly involve at least seeing you do it, whatever else it might involve.

But, even then, it is in such cases that seeing-as comes into play in a new way, different from the phenomenon Wittgenstein considered, and one that throws light on what pretense and going-through-the-motions are on this level of molecular actions. More about this presently.

To consolidate gains made so far, let us look at the upshot: the parts of a molecular action are themselves actions (atomic), things that the agent *does*. (He *walks* to *spy*.) An atomic action also has parts but, since these are bodily movements, they are not "proper" parts of the action as such. (Think of how this helps Kant.) The arm rising with flexed muscles is not, *as a bodily movement*, a proper part of, say, the action of opening the window. As Paul Ziff says, it is a "component of behavior," not itself an instance of behavior.[9] But this (atomic) action of opening the window *is* a proper part of the (molecular) action of letting fresh air in. One *does* open the window in *doing* the ventilating. But doesn't one raise his arm in opening, or to open, the window? Isn't that also something the agent does? Does not he raise the window with his arms? There is a temptation to say he does and that therefore raising the arm is also an action of his, and at this point the notion of action seems to fuse with that of bodily movement.

But see what happens if we pursue this line of thinking and talking to the bitter end. In this vein, we would ask next, what action does one take to raise his arm? He flexes his muscles. And to flex his muscles? He triggers the appropriate nerves. And how does he do this? By doing something with a part of his brain. All these are actions that the agent takes or performs. He *does* all these things. But how on earth does he spark his brain at just the right place, or with what does he do this? He does it with an act of will, guided by intelligence—roughly, with his mind, acting on the pineal gland at the center of the brain. Thus are we put back three centuries in our thinking and talking, in this vein. What we need now, to outlaw this sort of recession into the brain and finally into an "inner" mind, is a field of theory of persons-in-action, a field that we lapse out of when we begin thinking and talking in terms of physics, physiology, and neurology. What these report are movements, not actions. Thus, though it is true that one *moves* this way and that, as he *acts* in ways that involve such movements, it is a logical mistake to suppose that these movements are things he *does*—actions he performs. Otherwise we must think of a man as *doing* ten thousand things he is not even conscious of, and this in effect makes most—if not all—of what we are alleged to be doing lose the name of action. But we need also a *unified* field theory that makes intelligible the (logical) connection of doing or "acting" with "moving," and this is indeed a tricky business.

---

[9] "About Behaviorism," *Analysis*, vol. 18 (1957–58), pp. 132–136.

Ordinarily, if someone found your window open and asked you, "What did you open it with?" the circumstances that would give this question a use would be, say, the window's being too high to reach without a stick, or heavier than usual, or previously stuck, etc. In such circumstances, it would be perfectly in order to answer "Nothing," if you did not use a stick, etc. You didn't use anything to implement the action. You could quite consistently add, "I did it with my bare hands," giving this as a reason for answering, "Nothing." Your hands, your arms and muscles, are parts of you-in-action, such that *nothing* intervened between you and the window, nothing was used, in the atomic action of opening it. Ryle might have noticed the appropriateness of answering, "Nothing," even at such junctures.

And if the window is of the usual manipulable domestic type in good condition, and is found open, nobody would have any occasion to ask what you opened it with? Not because they already know the answer, but because the question would be meaningless.

Of course, raising your arm and flexing your bicep would be an action, something you do, were you to put it on exhibit for, say, your girl friend, to gain her confidence in you as a body-guard. Then you would be doing something *to* your arm as previously you did to the window. But this would be a clearer-cut case of senselessness for your companion to ask you what you did that with. She might well ask you *why* you did it, and the answer would be a reason for the *action*, not a reference to a cause instrumental to the movement. (Rube Goldberg's cartoons obliquely caricature the idea that an elaborate concatenation of instruments is to be used in the performance of a simple atomic action.)

With this detour behind me, detailing what went before, I am ready to focus on pretense, and on seeing-as in the new sense.

I cannot pretend to open the window—to "be opening" it is something else (Austin)—as I cannot pretend to walk. These are atomic actions, and perceptible. You may, without doubt, see me do them. But my ventilating the room may be a pretense, even if the room gets ventilated as an effect of my opening the window. Opening the window would be an instrumental part of the molecular action of eavesdropping, not of ventilating the room, if there were lovers outside whose private homilies of love I wanted to overhear. In that case, the ventilation would not be anything I did, *my action*, even if it occurred because of something I did do—opening the window. I would pretend that I was ventilating the room, to camouflage what I really was doing, eavesdropping. Thus would I only be going through the motions of ventilating the room.

It is in such more complex cases of at least two actions being performed—opening the window and eavesdropping in this case, the latter molecular action including the former as a means—that a serious question about perceptibility arises. You may fail to see me eavesdropping even though you are watching me, as you cannot fail to see me open the window. And what veils what I am really doing would be my pretending to be doing another molecular action, ventilating the room that would also require me to open the window.

It may be objected that ventilating the room was something I really did—it was my action—despite my only pretending to do it as I eavesdropped. And the reason for this objection is that were I actually ventilating the room, that is, not pretending to do it, I would, in the circumstances, do exactly what I did as an eavesdropper. This is why, so the contention goes, that the pattern of "bodily movements" may be "seen as" either the eavesdropping or the ventilating, with impunity; the implications being that the decision as to what really is going on does *not* rest with perception after all. There is always the *purpose* or intention that is out of sight; and this has to be inferred from what is seen and is what determines what molecular action is actually being performed. So again we run into the picture of an inner somewhat that is imperceptible—the invisible machinations of the mind behind the simple (atomic) actions that are perceptible thanks to their approximation to bodily movements.

But this is a mistake. Such an argument overlooks what Wittgenstein called "fine shades of behavior," And it is precisely the noticing of these that is realized with the help of a supposition or interpretation, something "in mind," that converts this sort of seeing-as into a revelation of what the agent really is doing where this is a molecular action, the composition of the fine shades of behavior into one action. This is the idea that I want to clarify, in conclusion.

We have first to attend to the ambiguity of "behavior" in Wittgenstein's phrase (and Hamlyn's). This could, out of context, mean a set of motions of bodily parts. But in speaking of fine shades of behavior, Wittgenstein meant the great variety of little things that *a person does* in the performance of more complex and complete actions. To notice the agent doing these elementary atomic actions in context, is, to "see" the agent "as" performing the molecular action that *actually* engages him. But such noticing and seeing-as is feasible only with the thought in mind that such-and-such molecular action is being performed, not the pretended one; it is as if, in such interpretive looking at what is going on, the supposition or thought "unfolds" into the perception of the molecular action. (Was it Strawson who said something like this?) A similar unfolding occurs in the perception of the duck-rabbit picture's aspects, though with the crucial difference noted above. If you think rabbit as you look, you will see rabbit. Think duck, and that is what you see. But, nevertheless, the result is a *perceptual* confrontation.

You and I are seated in the room, talking. There is a fine summer moon outside. You know that Harry and Helen are enjoying it nearby in the garden. You complain that it is warm and stuffy, and I see you get up and open the window, then move your chair closer to it. This much you might do, whether you really were ventilating the room or eavesdropping. But I can *see* which of these actions you perform if I look with that question in mind. It is just this apperceptive stance that shows me the fine shades of behavior the constellation of which atomic actions I see as the molecular action. The latter practical "aspect" dawns on *me*—thanks to my way of looking—or on the *figure;* it doesn't matter which of these alternative expressions is used, since the appearance of the aspect and the way of looking presuppose one another. If you are generally a good conversationalist and I know you have a live interest in the topic of our conversation, and you are not as attentive as usual, and you open the window *quietly* and through it I see the lovers close by on the lawn in the moonlight, it will pretty soon dawn on me what you are up to. And with this supposition in mind, I shall notice further the tilt of your head that cocks an ear at the window, while you absently continue your talk with me; and the like. All these atomic actions are the fine shades of behavior in the composition of which I *see* you eavesdropping, and that you are only pretending to ventilate the room, going through the motions only.

What differentiates the experience of such aspects from that of, say, the rabbit-aspect of the picture is that *its "base" is already a person, a man, engaged in the fine atomic actions I simply see him doing* and that jell into the pattern "seen as" the molecular action the agent is *actually* performing. Such seeing-as amounts to seeing what people are doing, or the molecular actions comprising their conduct. This *is* the perception of molecular human actions, and it is such that concern us most of the time.

Now, if the base-figure consisted of physical movements instead of atomic actions, there would be a chasm to be crossed by inference to actions of *any* sort; and this would leave room also for the ghost of the more fundamental scepticism about the very existence of other minds or persons. It is this bogus notion that we see "bodily movements" and must somehow see *these* "as actions" that leaves the old question standing about "logically adequate criteria" for the inference to actions, or about evidence for the seeing-as that claims to reveal the action on a base of physical movements. Strawson has, by clear intention more than by his explicit treatment, done much to lay this ghost. The way he couches his argument, however, leaves the ghost of this ghost—a second-order ghost—still haunting him. So he continues talking in effect about evidence for the actions we see, camouflaging "evidence" behind the phrase "logically adequate criteria."

My suggestion is that, as you do not infer that your friend is smiling at you from the observed changing physical contours of mouth and cheek as physical objects—a robot or a Punch-and-Judy-type doll could be made to go through such motions—neither do you, *in some situations*, infer the warmth and friendliness shown in the simple atomic action of smiling. There, you see your friend smiling *and* you see this "as" his show of affectionate confidence, meaning to inspire confidence in you. His smiling (atomic action) is not "evidence" in any sense for his molecular action of putting his friendship on exhibit, as the duck-figure is not evidence for the duck-aspect that visibly animates it. The immediate entourage of his fine shades of behavior is the condition of their jelling into the perceptible molecular action. This is what the perception of people by people, of persons by persons, amounts to. Any reduction of such perception of human beings in action to observation of physical movements, with its inevitable consequences of having to grope in the dark for evidence or criteria for what *actions* are going on, is the old mistake of trying to foist the model of physical object observation on the perception—I like to call it "prehension"—of persons by persons. So firmly does the latter stand out in its own right that, if you were asked to describe the physical components you "observed" in making the alleged conceptual jump to the action of your friend smiling at you and what it expressed, you would find that you had overlooked those movements in favor of perception of what your friend was *doing*.

Of course, in other situations, you may fail to see, you may be unable to see, what your companion is actually doing, even when he is not pretending, and thus be left in an agony of doubt. That he smiles at you you cannot doubt (the atomic action). But whether, in doing this, he is or is not acting "on principle," in a sort of calculating condescension, may be practically undecidable either by perception or further inquiry involving thought and inference. This is the tragic predicament of, for example, the black and white people in the United States. They simply cannot, on the whole, *see* what they are doing, the one the other. James Baldwin's *Another Country* is a fine, dark picture of this terrible predicament. It is the more frightening because, in such matters, the decision, if any, rests finally with perception.

Summarily, I have argued that perception by a person of what another is doing amounts, in the usual case, to perceiving a pattern of "fine shades of behavior," each itself an (atomic) *action* unmistakably present, not just an "observed movement." It is on this pattern that the molecular (intentional) action dawns as "practical" aspect—what the pattern is seen as, and such ("prehensive") seeing amounts to seeing what the person is "really" up to. This concept of molecular action could be replaced by the concept simply of the *intention* that perceptibly animates the pattern of elementary actions, itself not another action at all. But this revision would not upset the thesis about perception being decisive, even where the question of "criteria" is unanswerable (because irrelevant). When what is in question is seen, there is not (logical) room enough for inference.[10]

University of North Carolina
at Chapel Hill

Received July 1, 1966

---

[10] This point is developed in my "Too Obvious for Words," *op. cit.*

American Philosophical Quarterly
Volume 4, Number 3, July 1967

IV. COLORS AND SENSATIONS: OR HOW TO DEFINE A PAIN OSTENSIVELY

CHARLES B. DANIELS*

WHEN a person learns the color words, he learns at least three things which may not be unrelated:

First, he learns to see sameness and difference of shades of color.[1] He must be able to get what the teacher is talking about when the teacher points at various samples and says "That is (is not) the same shade of color as that."

Secondly, he must master whatever it is that orders shades of color, that makes them similar. This is not simply a matter of having seen samples of *all* shades and remembering which ones are called "red," which "blue," and so forth. No one has ever seen samples of every shade of color. Perhaps it is even logically impossible, inasmuch as it might be necessarily true that someone could always see a difference of shade where no one had seen one before. What I mean by the "order" among shades of color will be understood if one imagines oneself facing three samples of red of different shades placed side by side. The samples can be placed in different arrays left to right. In some, if we imagine the outside samples being blended into the center sample, the same shade of color will appear on both sides of the center. In other arrays it will not. In those in which it does not, the three samples are *in order*.

Thirdly, he must learn which shades are classed as being of the same color. The teacher may say that all shades of color like that of a given sample, or between that and the shade of another sample, are called by the same name, say, "red."

Color, sameness of shade of color, and similarity of shade of color are things that can, with certain reservations, only be seen. People who are color-blind, when they complain that they cannot see what it is that is the same, similar, or different about the things the teacher points at, will be told "You either see it or you don't. I can't *tell* you how to spot these things." But they are quite capable of using the color words in making statements. Like the totally blind, the problem for them is that their knowledge of the truth of statements about the colors of certain things owes itself wholly to contingencies like that of the bottom traffic light remaining green. Remaining green is neither a logically necessary nor a logically sufficient condition of remaining a bottom traffic light. If the color of the light were to change, and no one told those who are color-blind what the new color was, all they could do is to make mistakes or guesses. Their knowledge claims about the colors of things, provided, that is, the things they are talking about are not by definition of one or another color, are wholly dependent upon the word of others. From this it can be seen that the teacher of the color words can be corrected by his pupil on points of sameness or similarity of shade. Not all of us can see these things equally well.

Colors are ostensively defined. The teacher points to an instance of a certain color and says to the pupil, "That is an instance of the color, red," or something like that. The teacher knows that the sample is an instance of red because he can see that it is, and if the pupil is not color-blind he will see what the teacher is talking about too. It is possible for both the teacher and the pupil to see *the same instance* of the color at the same time. Of course, if the pupil asks "What do you mean, 'color'?", the teacher will have to find out whether he sees sameness and similarity of color. If he doesn't see it, the teacher cannot help him by sitting him down and telling him what it is.

Sensations, I shall argue, can also be ostensively defined. By sensations I do not mean emotions or other things which have objects, as when we feel pain at someone's departure; nor do I mean the kind of thing we are talking about when we say "He's just a pain." What I do mean will become clear as I proceed.

---

* I wish to acknowledge a debt to Sam Coval and David Pears for many enlightening conversations about these matters and for having unburdened me of many really badly mistaken beliefs—without, of course, pinning upon them some of the wilder ones I here espouse.

[1] I am talking loosely here and will continue to do so. If the reader objects he may take me to be talking about shades or degrees of any one of the three dimensions of color, what artists call hue, brilliance, and saturation.

How can we define sensations ostensively? The answer is simple: The teacher indicates an instance of a certain sensation and says to his pupil, "That is an instance of the sensation, pain," or words to that effect. Frequently, though, people are puzzled as to how the point gets across. After all, what happens is that, say, the teacher gives the pupil a hard pinch and says "That is an instance of the sensation, pain." The pupil jumps, he winces, he cries out, and perhaps his eyes water. Why isn't the the sensation, pain, taken by the pupil to be the pinch, the jump, the wince, the cry, and the moist eyes, *and nothing more?* What the teacher can see and hear seems to be all the teacher can know of the pupil's pain, and so how can the pupil *take him* to be talking about anything else? Why isn't the pupil turned into a little behaviorist?

Well, where we go wrong is by forgetting that almost without exception in a situation like the one I have described it is also the case that *it hurts.* The pupil doesn't really have to see anything or be in a position to see anything to feel what the teacher is talking about. Of course, if the pupil who has been pinched asks "What do you mean, 'sensation'?", the teacher will have to find out whether the pupil actually feels sensations. If the pupil just doesn't feel sensations, the teacher is thwarted. One either feels them or one doesn't, and that seems to be the end of the matter, just as in the case of colors where one either saw them or one didn't.

Pointing is not the *only* way that one person can indicate something to another. In teaching colors, the teacher might take the pupil's head in his hands and turn it forcibly toward the sample he wanted him to see rather than pointing at it. Then, too, let us remind ourselves of the many different ways we have of indicating what we can hear to one another, among which pointing is not one that figures prominently. Now one method of indicating something that people can *feel*—more specifically, let us say, a pain—is by pinching them hard, though no one will feel pain when anesthetized, just as no one will see color when color-blinded.

One way the teacher can know that what the pupil sees is indeed an instance of the color he happens to be teaching him the name of is by indicating something that he can see at the same time himself. What the teacher and pupil both see is one and the same instance of color. It is not, however, essential that the teacher cite as instances things that both he and the pupil see simultaneously. He might say, "Most lemons and your living-room walls are (are not) the same shade of color."

It would make the concepts of the various sensations somewhat less of a gigantic mystery if we could show that the same thing takes place in instructing someone in the use of sensation words, *or could take place.* Can we say that one way the teacher can know that what the pupil feels is indeed an instance of the sensation he happens to be teaching him the name of is by indicating something that he can at the same time feel himself? Can what they both feel be one and the same instance of a sensation? The problem really seems to amount to this: what is it for something to be an instance of a certain color or sensation?

One thing the reader should bear in mind throughout is the fact that pain hurts. This itself might serve to explain why we usually do not teach people what pains are by giving them pains, and check on what they feel by giving ourselves pains. To do this one would have to be both sadist and masochist.

Instances of colors are identified by virtue of their being things *besides* instances of colors, things which have spatial location (like crocodiles, pigs, etc.), or because they happen to be parts of things like these, or simply by virtue of the fact that they are at least to be found *somewhere*—like the blue above on a cloudless day. "What did you say was an instance of the color, red?", "Why, what we just saw when we looked at the screen!"

Instances of colors are only *re*identified "by courtesy," so to speak, inasmuch as the things they also happen to be besides instances of colors are reidentified. The Russian flag we see is the same color as it was yesterday. But only because the Russian flag we see is numerically the same flag as the one we saw yesterday might we say that the Russian flag we see is numerically the same instance of red as the one we saw yesterday. Few people are interested in *re*identifying instances of colors anyway, at least over more than minimal periods of time. This "as-you-will"-ness attends even the simple identification of instances. If we stand back to back and elevate our gaze slightly, do we see *the same instance* of blue? As you will.

Instances of the sensation of pain (of warmth) may be identified by their spatial locations. "What did you say was an instance of the sensation of pain (warmth)?", "Why, what we just felt when we touched the electric fence (got close to the fire)!" or "what we just felt when we got our hands pinched in our friend's car door!"

It may be objected here that the teacher and the pupil do not feel the same *instance* of pain because each feels the pain only in his own hand and not in the other's hand. Where the teacher's hand was is where the teacher felt the pain. The teacher did not feel the pain in the pupil's hand. The teacher's hand never overlaps the pupil's hand spatially. Thus, since instances of pain are identified by spatial location, the teacher and the pupil actually feel different instances of pain. Indeed, no two people feel the same instances of pain because no two people overlap spatially, save Siamese twins, of course.

This objection can be met by noting the distinction between *where* one feels an instance of pain and *what* one feels an instance of pain *in*. There are, in the analogous case of color, at least two ways to answer the question "Where did you see the instance of red?", a fact which indicates that there are at least two ways to ask it. The first way is by saying something like "on the screen." Another way, however, is by saying "out of the corner of my eye." Now when a teacher sees the same instance of red as the pupil sees, he certainly never sees it out of the pupil's eye, much less out of the corner of the pupil's eye; nor do we require that he do so to see what we term the "same" instance of red.

Feeling a pain in one's hand may be likened to seeing a color out of the corner of one's eye. What would correspond to seeing an instance of red on a screen? Feeling an instance of pain in someone's car door, or along an electric fence. If one is going to ask about the whereabouts of the instance of pain in this latter sense, it is where the teacher and pupil both feel it—in the car door or along the fence.

Yet how are we to explain the fact that people rarely seem to give this kind of answer to the question of the whereabouts of a sensation? Perhaps this can be explained by citing one difference that exists between sight and the sense of feeling: the location of the colors we see may be separated from our eyes by miles; the location of the sensations we feel must be contiguous to our sensing organs. The answer to the question that asks the location of the thing seen is not often given by saying "Out of the corner of my eye." This answer is usually of little help in telling us the location of the thing seen; but more important we can just say where its location is in the other way. But the answer to a sensation location question, because of what I believe to be a necessary contiguity between the sensation location and the location of the sensing organ, what one feels it in, does give the whereabouts of the sensation. An answer to one of these questions gives the answer to the other, so asking them separately is really of little point.

This last contention may be supported further by the case of the Russian woman who was reputed to be able to feel colors. Why is it that the press chose to describe her case as one of "feeling colors with her fingers," rather than one of "seeing colors with her fingers"? It may be this: if she were to be said to see colors with her fingers, she would have to be able to insert her fingers through the crack of a narrowly opened door and report the colors of the furniture in the adjoining room without peeking with her eyes or coming into contact with the objects she was talking about. If she could do this when and only when the light was on and her fingers were in a suitable position in relation to the objects whose colors she was describing, then we might well say that she was seeing with her fingers. One might distinguish further between seeing colors with one's fingers and hearing colors with them by tests with light, opaque objects, etc. But when she can only tell us the colors of objects *by coming into contact with them*, her case is described as one of feeling colors. Because we know where she feels them (what she feels them with or in), we know where they are.

It might be questioned here that the locution "feeling colors" makes sense, despite its use by the press and its apparent intelligibility to readers. My personal belief is that it does make sense. If the question were raised "What would it be like to feel colors?", on the model of the question "What would it be like to have a circle with corners?", I would have to answer that I don't know what it would be like, because I cannot feel colors. Nor would anyone understand what it would be like in this sense who couldn't do it. But nor do we know what it would be like to raise our hair without taking hold of it. It seems to me that we would be in the position of a blind man with respect to colors, or a man who has never had use of his arms with respect to raising them. My argument, though, does not depend on the intelligibility of the phrase "feeling colors." It is rather that, even granting that the phrase is incoherent, people used it because they regard feeling as a *non-distance* sense with regard to "simples," like colors, pains, etc., and they would regard it as *more incoherent* to say that the woman *saw* colors with her fingers, or heard them, if to discern them she had to make contact.

Many readers might also, no doubt, want to object to what I have been saying by pointing out that we would say in the teacher-pupil case that what they each felt was the same *kind* of sensation, e.g., pain. This is not, however, feeling the same instance of pain.

Just as there are indefinitely many shades or degrees of hues, brilliances, and saturations, there are indefinitely many shades of sensations of different kinds, a wide range of which happen to be classified as "pain." There is then the further division within what is called "pain" into different, often non-exclusive sub-kinds of pain, such as twinges, pangs, stitches, throbs, gnawing, burning, shooting pains, cramps, stinging sensations, shocks, and aches of all sorts, headaches, backaches, etc. It is doubtless true that both teacher and pupil feel the same kind of pain in some manner of kind. To say this, however, does not controvert my assertion that they feel the same instance of pain. If they feel the same instance of pain, then of course they feel the same kind of pain in some manner of kind.

In this way one learns the names of the various sensations. How does the teacher know that what he is indicating is an instance of the sensation he wants to teach the name of? Because he can feel it himself. If the pupil asks "What do you mean, 'sensation'?" and asks it seriously, then he may be "sensation-blind," perhaps partially, perhaps completely. He may, however, be more sensitive to sensation differences than his teacher. He may be able, by sensations alone, to tell differences in the gauges of the needles he is being pricked with, something that no one else in the world can do. To learn the sensation words, then, a person will have to learn what sameness and similarity of sensation are, and then what they are called.

Thus far my explanation has been drawn from an analogy with the explanation of color words. There is a difference, though, in that while colors can exist unseen, sensations are usually thought not to exist unfelt. This might be put in terms of ontological dependence. Take the terms "crocodile," "head," "bottom," and "shadow." There are no crocodiles that are necessarily crocodiles of anything; we do not have to be able to identify another individual before we can identify a particular crocodile. Now while there may be heads which are not the heads of anything—perhaps the Bodley Heads provide examples—it is hard to imagine coming across a shadow that is not the shadow of something or other (even a five o'clock one), or a bottom that is not the bottom of something or other. Suppose we find what looks like the bottom of a glass jar, but proves to be a naturally formed piece of quartz. What we have found is then not a bottom, although it looks like one.

The contention is that sensations are like shadows and bottoms. The latter can be seen, the former felt. Two people can feel one pain, just as they can see one shadow or bottom. Shadows, however, can only be (or only are) seen, and in the same way sensations can be (or are) only felt.

There is a disanalogy here, too, to be sure. Shadows need not be seen by the individuals whose shadows they are. The necessary relation that holds between a person and his shadow is not the relation of seeing. The necessary relation that holds between a sensation and its owner is said to be precisely that of feeling.

To summarize the position I have sketched out thus far: sensations are shareable, they are located in space, they are ontologically dependent upon their owner or owners, and this relation of dependence is thought to be that of *being felt*.[2]

Nonetheless, there is a further *and separate* matter of whether (1) if a person feels (or has) a sensation, he can be mistaken about it, and (2) if a person feels (or has) a sensation, he can be mistaken about what kind of sensation it is. I myself think that both kinds of mistakes can be made, but it would be much too long a project to argue it here. I shall just attempt a description of the latter kind of mistake.

Imagine a torture situation: prisoner strapped face down to table. Information to be extracted. Iron being heated to red-hot before eyes of terrified prisoner. The torturer takes up iron, brandishes it before the prisoner's face, walks behind him, and, completely unbeknownst to the prisoner, applies an ice cube to the prisoner's back. He screams, faints.

Why does the prisoner scream and faint? Because he mistakenly thinks that he feels a severe pain, when in fact the sensation he feels is not that at all.

---

[2] One might question this. Perhaps we have many unfelt sensations as we go through our days, in our feet when we walk, in our fingers when we write, etc. Perhaps the necessary relation is not that of *being felt*. I shall not, however, insist upon this point.

The inquisitor knows what it is because he feels in his fingers what the prisoner really felt in his back.

Or imagine the reverse. Prisoner in same position. But when the inquisitor turns away to heat the iron, the warder winks conspiratorially at the prisoner and guardedly indicates something behind the prisoner's back. The prisoner cranes and catches a glimpse of a bowlful of ice cubes. Then the warder finds an opportunity to whisper in the prisoner's ear "He'll use an ice cube, not the iron. We've got orders not to hurt you. Don't be afraid, just play along. Scream and pretend to faint." The inquisitor (who is really in cahoots with the warder) then takes up the iron, brandishes it realistically before his victim's face, walks behind him, quietly rattles the ice cubes, and touches the red-hot iron to the prisoner's back. Prisoner follows instructions by screaming and pretending to faint. Thirty seconds pass (or as few as one needs to make my story credible) and then the inquisitor says "He's fainted, damn it! Hand me that towel and I'll wipe the water off his back so he won't realize what's happened. We've got to have another go at making him talk, you know" and proceeds to apply a strong local anesthetic.

Here again the prisoner makes a mistake. He feels a severe pain and mistakenly thinks it is just a sensation of coolness. In this case, however, the inquisitor isn't masochistic enough to check. In both cases the prisoner does feel a sensation. He is simply mistaken as to the kind he thinks it is. Thus, I would make it possible for two people to feel the same sensation, and disagree if they were to state their own evaluations as to its kind.

It might be objected here that what the prisoner is mistaken about is not the sensation's kind, but its cause. To this I answer that he is mistaken about both. One can construct a case in which a very nervous prisoner knows that one of a number of things will be applied to his back—a red-hot poker, a lazer, a moray eel, an ice cube, a finger—but not which one. The ice cube is applied and he screams and faints because he mistakes the sensation for one of pain. He may also, but need not, judge as to its cause. Similarly if two savages unknowingly touched an electric fence, they would not mistake the sensation, but wouldn't have a clue as to its cause especially if the current were switched off immediately afterwards.

Yet how can this theory deal with sensations like toothaches, headaches, nausea, and giddiness that we feel well within our bodies? First, it could be said that with regard to some sensations "within," we answer questions about location in the "on the screen" manner, rather than in the "out of the corner of my eye" manner. In these cases we are notoriously bad at fixing the pain in organs, say, in the pancreas. We locate the sensation, and then by our vague general knowledge of anatomy, try as best we can to figure out what organ is the sensing organ. Of course if we *call* a pain a headache and it is one, this itself doesn't entail that someone else couldn't have it, or that if he had it, he would have to have it in *his* head. The term "headache" indicates the pain in the "out of the corner of my eye" manner, and someone else might perhaps have the very same pain in his finger.

If these sensations are located somewhere in the space enclosed by the epidermis, however, it seems that the only way that another person could, *qua* masochist, feel the same instances we feel is by inserting one of his sensing organs into our bodies and "focusing" it in the right way. What would be the right way to "focus" an organ in such a case? The fact is that rarely if ever do we ourselves know how our own organs are "focused" in such cases, which is not to say, however, that we have no idea of how to set about finding out.

The analogy here is between the use of spectacles, microscopes, mirrors, etc., to aid our eyes to "focus" upon visible objects. Similarly we might, if sadistically inclined, use shocks, red-hot pokers, etc., to enable our pupils to "focus" upon sensations. In the way in which we are ignorant of causal aids that would enable some persons to focus upon instances of colors, so we are ignorant of ways that might enable interested parties to "focus" upon our internal sensations in some cases—we simply don't know how we do it ourselves. But then what would count as being "the same instance of a sensation" might itself become clearer—and perhaps other than I have pictured it—as we find these things out.

As to other sensations, like giddiness, why shouldn't we count what we both just felt when we were on the tilt-a-whirl as the same instance of giddiness?

There is another objection to consider: We do talk about redly colored objects; we do not talk about painfully sensationed objects. But then, of course, we talk about painfully itchy sensations and not about itchily painful sensations. Yet how much weight are we to attach to this usage? Can't we imagine a language in which people do say things about what they term interchangeably "itchily painful sensations" and "painfully itchy sensations"? also about "painfully sensationed objects"—perhaps thumbscrews, or pinching shoes? Why not!

By way of conclusion, it might be helpful if I were to sketch out a view that contrasts with the one I have been putting forward. My contention has been that the "feel" in statements like "I feel a pain" can be taken to be transitive in a metaphysical sense. The pain I feel is an object of sorts. Another view, though, is that upon analysis the entire phrase "feel a pain" says something that is at bottom *in*transitive, as does the phrase "sitting quietly" in the statement "I am sitting quietly." There is no object of any sort. If one favors this as the more true-to-life account of sensations, then one will not say that two people can feel the same instance of a sensation. My sitting quietly and your sitting quietly will be two different instances of the same kind of thing, i.e., sitting quietly. Again under this interpretation there will be a further and separate matter of whether: (1) if a person instantiates feeling-a-pain, he can be mistaken about his instantiating feeling-a-sensation, and (2) if a person instantiates feeling-a-pain, he can be mistaken about the kind that "adverbially" qualifies his feeling-a-sensation, i.e., *-a-pain*, here taken to parallel *quietly* as a kind that qualifies sitting.

I think my torture cases will apply here too, although the explanation of how the inquisitor knows what is the case will differ. But let us suppose the opposite, that these mistakes cannot be made, that if a person feels-a-pain, he knows that he feels-a-pain. We might, then, call this kind of knowing "direct" or "immediate" knowledge. Just to rid ourselves of possible prejudices, let us say that the predicate "sitting quietly" is of this kind, i.e., that if one is sitting quietly, then one knows that one is sitting quietly, rather than continuing to talk about feeling-a-pain.

As things are pictured here, it will be impossible for another person to know "directly" that I am sitting quietly, since his sitting quietly is a different instance of sitting quietly from mine, and what he knows "directly" will be that he is sitting quietly and not that I am. To the ways of knowing that normally come to mind, e.g., seeing, hearing, being told by a reliable source, etc., there will have to be added others, e.g., sitting quietly, feeling-a-pain, etc.

At this point confusion often arises. We are sometimes tempted to think that because I cannot know that another person is sitting quietly "directly," I am limited to knowing that he is sitting quietly by way of criteria, behavior, induction; or worse, that I am limited to evidence that can never be "sufficient" or "adequate" in some or any sense of these words. We may then even go on to make up other additional ways of knowing for certain things, for example, for Daniels' sitting quietly, being Daniels as he sits quietly. We then say, or more likely have already said, that you cannot know that I am sitting quietly in the way that I do, i.e., "directly." And, of course, this will be true, in fact true by definition.

Yet we must remember that our word "directly" was *not* introduced in opposition to "criteria," "behavior," "induction," or "evidence." The actual contrast was that between knowing that p merely by virtue of p's being the case, and not knowing that p merely by virtue of p's being the case. The latter subdivides into two cases: not knowing that p (period) and knowing that p not merely by virtue of p's being the case. And it is a further, and I think, unjustified step that equates knowing that p not merely by virtue of p's being the case with knowing that p by way of criteria, behavior, induction, or evidence. To be sure, if someone insists upon pre-empting the words "criteria," "behavior," "induction," or "evidence," and then makes the contrast, we can give him the words and still not be worried when he says that the only way we can know that he is sitting quietly is by way of criteria, or etc., because all he means is that we can't know that he is sitting quietly merely by virtue of its being the case that he is sitting quietly.[3] What remains for us is knowing that he is sitting quietly, not merely by virtue of its being the case that he is sitting quietly. And what is wrong with that?

"Bodies" are often mentioned in these contexts too, in that we are not supposed to be able to know things that are features of bodies "directly." But again, why not? Nothing in the notion of something of which it is necessarily true that if one has it, one knows one has it, makes it impossible for it to be a feature of someone's body, nor, so far as I can see, does anything in the meaning of the word "body" make it impossible for things of this sort to be features of them.

The general position, the one that by itself does not embody any of these particular mistakes, is that there are two kinds of facts in the world: the true proposition to the effect that someone knows one of the first kind of fact will entail that there is another different fact; the true proposition to the effect that someone knows one of the second kind of fact will not entail this. For example, if it is a fact that someone knows that I am six feet tall, then something else will have to be a fact besides the fact known: that the individual who knows this has, say, measured my height, or something of the sort. If, on the other hand, it is a fact that someone knows that I am in pain, then nothing else will have to be the case besides the fact known: the fact that I am in pain is the same fact as my knowing that I am in pain. But nothing has been said here about "criteria," "behavior," "induction," or "evidence" in the case of others knowing about my being in pain. They know that I am in pain, if they do, of course, not merely by virtue of its being a fact that I am in pain.

---

[3] It is interesting to note that Strawson, in his chapter on persons in *Individuals*, does use the phrase "behavioral criteria" to make just the contrast I have been discussing (that is, if I interpret him aright) and through the resulting misunderstandings (again, if I interpret him aright) has drawn more criticism than he warrants.

Yale University

Received April 19, 1966

American Philosophical Quarterly
Volume 4, Number 3, July 1967

V. MATHEMATICAL PROOF

JOHN L. POLLOCK

IT is almost a commonplace of contemporary philosophy that analyticity and *a priori* truth are to be explained in terms of "rules of language" or "linguistic conventions." But in this article I shall endeavor to show that this cannot possibly be true. I shall argue this by looking at a particular class of *a priori* truths—mathematical truths.

Consider that type of proof which is exemplified by ordinary working mathematicians doing mathematics who are not interested in the philosophy of mathematics and are not explicitly working within an axiom system. Let us call this "proof in the classical sense" or "classical proof." If we look closely at classical mathematical proof, it is apt to appear mysterious. The difficulty is, where does the mathematician get the basic principles from which his proofs begin? Where does he get his raw material? A proof does not begin from nothing. In any proof there are premisses which are not proved in that proof and from which the proof proceeds. Of course, in any given proof many of the premisses may have been proven previously by other proofs. But it can never be the case that *all* of the premisses used in *all* of these proofs have been proven. The mathematician had to start somewhere, and what he started with he did not prove. Where do these basic premisses come from?

A popular and influential answer to the above question has been that these basic premisses are axioms, laid down by stipulative conventions which give us the meaning or content of our mathematical concepts. On this view the basic premisses are man-made—mathematicians themselves (or perhaps the linguistic community as a whole) "decided" what to take as axioms. The axioms are established by convention, and then mathematical proof can be viewed as the mechanical development of these axioms.[1] For example, Carl Hempel, in his article "On the Nature of Mathematical Truth," writes:

> The rigorous development of a mathematical theory proceeds not simply from a set of definitions but rather from a set of non-definitional propositions which are not proved within the theory; these are the postulates or axioms of the theory. . . . If we call the postulates and definitions of an axiomatized theory the "stipulations" concerning the concepts of the theory, then we may say that the propositions of the arithmetic of the natural numbers are true by virtue of the stipulations which have been laid down initially for the arithmetical concepts. . . . It is possible to deduce, from the Peano postulates and the additional definitions of non-primitive terms, all those propositions which constitute the classical theory of arithmetic, algebra, and analysis; and it is to these propositions that I refer . . . as the propositions of mathematics.[2]

The above view can be stated roughly as the view that mathematical axioms are man-made—that mathematicians did not *discover* that they were true—they *made* them true by convention. The only alternative to this is to maintain that mathematical axioms are not just *taken* to be true—the mathematician somehow *comes to know* that they are true. He did not *make* them true—they were true all along. My concern in this paper is to show that the first alternative (conventionalism) cannot possibly be correct, and hence that the second alternative is the correct one. Then I shall have some comments to make about the way in which mathematicians come to know that their basic premisses are true. Thus my problem is primarily an epistemological one—what are our grounds for our knowledge of mathematical truths?

If this picture of mathematical proofs as the formal derivation of consequences from stipulative axioms is to explain classical mathematical proof, and take the mystery out of it, then we must also be able to mechanize the methods by which we develop the consequences of the axioms. In other words, we must be able to work within a particular system of formal logic. If we couldn't do this, then the proof that something is a consequence of the axioms would involve what is essentially a classical mathematical proof, and so could not without circularity be taken as explaining classical mathematical proof. Combining the mathematical axioms with the axioms and rules of inference of the system of formal logic, we get a formal axiomatic theory. It is then proposed that classical mathematical proofs are just somewhat sloppy versions of formal proofs which are part of the purely mechanical development of this formal axiomatic theory. Let us call this theory of mathematical proof the *Conventionalist Thesis.*

---

[1] I have not made this view very clear, but I don't think that its supporters have succeeded in making it any clearer. What does it mean to talk about stipulative conventions that no one ever overtly stipulated? The view under consideration has often been attacked on just this point. But because the view is somewhat vague, it is hard to know whether these attacks have really scored against it. Fortunately, I think that there are other grounds upon which it can be attacked much more conclusively, as I shall endeavor to show in this paper.

[2] *The American Mathematical Monthly*, vol. 52 (1945), pp. 543–556.

It has been my experience that the Conventionalist Thesis represents the predominant view of mathematical proof among logicians and mathematicians who, while not writing on the philosophy of mathematics, are nonetheless cognizant of it and are working in the foundations of mathematics, i.e., set theory, formal logic, and metamathematics. And I think there is indeed a strong *prima facie* case for the Conventionalist Thesis. "If classical mathematical proofs cannot be represented in this way," one may ask, "then where does the mathematician get the raw material from which he begins his proofs?" I think this is a very good question indeed, but the answer cannot be, "From stipulative axioms," as I shall endeavor to show.

Now let us look again at the precise content of the Conventional Thesis and see how it must be stated if it is to constitute a satisfactory explanation of mathematical proof.

First, there may well be different axiomatic theories in different areas of mathematics. For example, there may be one for number theory, another for set theory, another for the theory of complex variables, and so on. But we can combine the axioms of all of these theories into one set of axioms yielding a single axiomatic theory which, if the Conventionalist Thesis is true, is adequate for the whole of extant mathematics. Then to extend this theory would just be to extend extant mathematics. So let us suppose there is such an axiomatic theory T within which we can do all of extant mathematics. What can we require of T if it is to make the Conventionalist Thesis true?

The first thing that comes to mind is that if proof within T ("T-proof" for short) is to constitute an explanation of classical proof, then we must be able to tell that a T-proof is actually a T-proof without having to rely upon a classical proof to determine this. If we had to give a classical proof to show that a proposed T-proof is actually a T-proof, then it would be just as much of a mystery how we know that something is a T-proof as it is where we get our basic premisses in a classical proof, and so representing classical proofs as being just "sloppy" T-proofs would make them no less mysterious.

If we can tell that a T-proof is a T-proof without having to *prove* that it is, then we must be able to tell that it is just by looking at it. This means in particular that we must be able to tell by inspection that a line of the proof that purports to be an axiom really is an axiom. The only properties that a sentence can, simply by inspection, be seen to have are its formal or syntactical properties. Thus what makes a sentence an axiom must be some purely formal or syntactical feature of it. And as we can only check the sentence for finitely many syntactical features, there must be a finite list of syntactical features which are adequate to characterize all of the axioms of T. Otherwise we could not tell by inspection whether a sentence is an axiom. The axioms of T are supposed to be the result of stipulative conventions concerning what we are going to take as axioms. Thus the axioms must be given to us by a finite set of conventions each of which tells us that all sentences satisfying some particular syntactical condition are axioms. Note also that, as we must always be able to determine that an axiom is an axiom, given any axiom we must in principle be able to determine that it satisfies these syntactical conditions. If we now form the disjunction of the syntactical conditions specified by the conventions, we can say that the axioms of T are characterized as the set of all sentences satisfying this particular syntactical condition, and this condition must be such that, if a sentence satisfies it, then we can in principle determine that it does.[3]

Similarly, if any given line of a proposed T-proof is correctly inferrable from the previous lines of the proof using any given rule of inference of T, then we must be able to tell that it is simply by inspection. Otherwise we would again have to resort to a classical proof to determine that something is a T-proof. We can think of a rule of inference as telling us that if a given sequence π of sentences is already a T-proof, and the sentence ϕ bears a certain relation $\mathfrak{R}(\pi,\phi)$ to π, then the sequence π^* which results from appending ϕ to π is also a T-proof. This is the most general form of a rule of inference. As we must be able to tell by inspection that ϕ can be correctly inferred from π by a single application of this rule, $\mathfrak{R}(\pi,\phi)$ must be a purely syntactical relation, and furthermore it must be one such that if it holds between a given sentence ϕ and a given sequence π, we can in principle determine that it does.

---

[3] Note that any recursive or recursively enumerable set of axioms satisfies this weak restriction. But by stating it in this way rather than requiring that our axioms constitute a recursively enumerable set, which is the customary procedure, we avoid having to assume Church's Thesis.

Furthermore, as we must be able to tell just by looking at a T-proof that it is a T-proof, we must be able to tell by inspection that *there is* a rule of inference which allows us to introduce any given line of the proof. We can only check finitely many syntactical conditions $\mathfrak{R}$, so there must be only finitely many rules of inference in T.

Thus both the axioms and the rules of inference of T can be characterized purely syntactically. From this it follows that the notion of a T-proof can itself be defined in purely syntactical terms—a finite sequence of sentences is a T-proof if and only if each sentence in the sequence is either an axiom or else can be appended to the sequence of its predecessors using one of the rules of inference of T. And as we can always tell by inspection that an axiom is an axiom and that a sentence which can be correctly inferred at any given stage using the rules of inference can be so inferred, we can always tell by inspection that a T-proof is a T-proof.

This means also that theoremhood in T can be characterized syntactically. A sentence ϕ is a theorem of T if and only if there exists a finite sequence of sentences which satisfies the syntactical conditions necessary to make it a T-proof, and ϕ is the last member of the sequence. Furthermore, if a sentence is a theorem of T, then we can always establish that it is, simply by constructing a T-proof of it.

The theory T is supposed to contain all of extant mathematics. This means in particular that we must be able to formulate all sentences of elementary number theory (that is, all sentences of number theory that can be formulated in an applied first-order functional calculus in which the only non-logical symbols are individual constants referring to the natural numbers and symbols referring to '$+$' and '$\cdot$') within the language of T.

The theory P of Peano's arithmetic is the theory that results from combining Peano's axioms with first-order logic with identity. The theorems of P are certainly among extant mathematics, so it must follow that they are all theorems of T—that T contains P.

Using Gödel's methods, we can now enumerate the expressions of the language of T, letting $\#\phi$ be the Gödel number of an expression ϕ,[4] and then for any syntactical predicate $\mathfrak{F}$ of expressions of the language of T, we can construct within elementary number theory (and hence within the language of T) a corresponding number-theoretic predicate $F(x)$ such that we can prove classically that for any expression ϕ, $\mathfrak{F}(\phi)$ is true if and only if $F(\#\phi)$ is true. In particular, remembering that theoremhood in T can be characterized syntactically, we can construct a number-theoretic predicate $\mathfrak{P}(x)$ within the language of T such that we can prove classically that

(1) For any sentence ϕ of the language of T, $\mathfrak{P}(\#\phi)$ is true if and only if ϕ is a theorem of T.

We saw that if a sentence ϕ is a theorem of T, we can always establish that it is simply by constructing a T-proof of it. Then using (1) we can prove classically that $\mathfrak{P}(\#\phi)$. We are supposing that anything that we can prove classically can be proven within T, so it must follow that

(2) For any sentence ϕ of the language of T, if ϕ is a theorem of T, then $\mathfrak{P}(\#\phi)$ is also a theorem of T.

Thus the Conventionalist Thesis commits one to the following propositions:

---

[4] There do exist languages whose expressions cannot be Gödel numbered, and so it might be supposed that this provides a way of avoiding the consequences of this argument. Perhaps the language of T is such a language. But I don't think this actually provides a way out. There are only two ways in which a language may fail to be Gödel numberable: (1) it may have non-denumerably many primitive symbols (e.g., a primitive individual constant for every real number); (2) the syntax may be so irregular that the sentences of the language are not generated by any regular rules. But I think that both of these alternatives are precluded by the supposed conventional origin of the language of T. First, there must be conventions which generate the sentences of the language in some mechanical manner, thus making (2) false. And second, those conventions must start from a finite set of primitive symbols—no language that people could actually use could have infinitely many distinct primitive symbols (although it could of course have infinitely many expressions that are generated by concatenating the primitive symbols). For example English starts with the twenty-six letters of the alphabet and a few punctuation symbols, and then everything else is generated from these primitive symbols.

(CT) There is a theory T such that:

(i) There is a number-theoretic predicate $\mathfrak{P}(x)$ such that for any sentence ϕ of the language of T, if ϕ is a theorem of T then $\mathfrak{P}(\sharp\phi)$ is also a theorem of T;

(ii) we can formulate all of extant mathematics within the language of T;

(iii) a sentence of T is a theorem of T if and only if it is provable classically (and hence T contains P).

Some philosophers[5] seem to be of the opinion that it has already been shown that CT is false, simply on the basis of Gödel's incompleteness theorem. However, that is not the case. First of all, it is not clear that CT is strong enough to allow us to prove a version of Gödel's theorem about T. But even if it is the theorem will only say the following:

> Given any consistent theory T containing P and for which there is a number-theoretic predicate $\mathfrak{P}(x)$ such that for any sentence ϕ of the language of T, if ϕ is a theorem of T then $\mathfrak{P}(\sharp\phi)$ is also a theorem of T, there is a sentence of the language of T such that neither it nor its negation is a theorem of T.

But this does not yet establish a difference between classical proof and T-proof, because it might well be the case that there are sentences of number theory (Gödel's undecidable sentence among them) that are undecidable classically as well as within T. In fact, this conclusion is drawn explicitly by Hempel in his article "On the Nature of Mathematical Truth."[6] Assuming the Conventionalist Thesis, this would just be taken as showing that to make further progress in mathematics, we must adopt some additional axioms. Thus Gödel's theorem does not yet establish that there is a difference between classical proof and T-proof, and so does not refute CT.

To refute the Conventionalist Thesis we must draw out its implications still further. The thing to note is that if the Conventionalist Thesis is true, then we must be able to prove classically that all theorems of T are true. This is for the following reason: According to the Conventionalist Thesis, T-proofs are just more rigorous versions of classical proofs: they make the axioms and rules of inference used in the classical proofs explicit. Thus the rules of inference of T are rules that would be accepted as valid in a classical proof, and as there are only finitely many of them we can verify that this is so simply by inspection. So we can prove classically that whatever is proven from true premisses using the rules of inference of T is also true. Thus if we can prove classically that all of the axioms of T are true, it will follow that we can prove classically that all of the theorems of T are true. The axioms of T are among those principles that are accepted as true without proof and taken as basic premisses in classical proofs. Thus we can verify by inspection that each axiom individually is true. However, there may be infinitely many axioms, so we cannot simply examine them all to see that they are *all* true. But the axioms of T are supposed to result from finitely many stipulative conventions to the effect that all sentences satisfying particular syntactical conditions are axioms. These conventions are supposed to represent the actual source of our unproved basic principles in classical proofs, and it is by virtue of the fact that a sentence satisfies one of these syntactical conditions that we recognize it to be true without proof. Each of these conventions has the form: "Henceforth, each sentence ϕ which satisfies the syntactical condition $\mathfrak{J}(\phi)$ is true." But if we have adopted this convention (as it is supposed that we have classically) then we know, not just that each sentence ϕ, taken individually, which satisfies the condition $\mathfrak{J}(\phi)$, is true, but also that *all* such sentences are true. And as there are only finitely many of these syntactical conditions, we can survey them one at a time and see classically that all of the axioms of T are true. Thus we can prove classically:

(3) For any sentence ϕ of the language of T, if ϕ is a theorem of T then ϕ is true.

Combining (3) with (1) we see that

(4) For any sentence ϕ of the language of T, if $\mathfrak{P}(\sharp\phi)$ is true then ϕ is true, i.e., $\mathfrak{P}(\sharp\phi) \supset \phi$ is true.

Thus for any sentence ϕ, we can prove classically that $\mathfrak{P}(\sharp\phi) \supset \phi$ is true. This latter is also a sentence of the language of T, so according to CT it must be a theorem of T.

Let us further suppose that classical probability is consistent. If it weren't, then there would be no point in trying to reconstruct it formally. On this supposition, it follows from CT that T is also consistent. Thus the Conventional Thesis implies:

---

[5] Paul Bernays, "On Platonism in Mathematics," *Philosophy of Mathematics*, Paul Benacerraf and Hilary Putnam (eds.) (Englewood Cliffs, N.J.; Prentice-Hall, 1964), pp. 274–286, tr. from the French "Sur le Platonisme dans les Mathematiques," *L'Enseignement Mathematique*, vol. 34 (1935).

[6] *Op. cit.*

(CT^*) There is a consistent theory T and a number-theoretic predicate $\mathfrak{P}(x)$ such that:

(i) T contains P;
(ii) for any sentence ϕ of the language of T, if ϕ is a theorem of T, then $\mathfrak{P}(\sharp\phi)$ is also a theorem of T;
(iii) every sentence of the form $\mathfrak{P}(\sharp\phi) \supset \phi$ is a theorem of T.

Although Gödel's theorem was not strong enough to refute CT, we can now prove a theorem that is very closely related to Gödel's theorem which is sufficient to refute CT^*. This theorem makes use of the same methods and constructions as Gödel's theorem, but it has a stronger hypothesis and hence a stronger conclusion. The theorem is as follows:

THEOREM: If T is a theory and $\mathfrak{P}(x)$ is a number-theoretic predicate such that:

(i) T contains P;
(ii) if ϕ is a theorem of T then $\mathfrak{P}(\sharp\phi)$ is a theorem of T;
(iii) T contains the theorem scheme $\mathfrak{P}(\sharp\phi) \supset \phi$;

then T is inconsistent.[7]

Proof: Suppose T and $\mathfrak{P}(x)$ satisfy hypotheses (i)–(iii) of the theorem. Using Gödel's methods, we can find a sentence θ such that $\theta \equiv \mathfrak{P}(\sharp{\sim}\theta)$ is a theorem of P, and hence of T.[8] So,

(5) $\theta \equiv \mathfrak{P}(\sharp{\sim}\theta)$ is a theorem of T.

By clause (iii) of the hypothesis of the theorem, $\mathfrak{P}(\sharp{\sim}\theta) \supset {\sim}\theta$ is a theorem of T. Thus $(\theta \supset {\sim}\theta)$ is a theorem of T, and so

(6) ${\sim}\theta$ is a theorem of T.

Then by clause (ii) of the hypothesis of the theorem, $\mathfrak{P}(\sharp{\sim}\theta)$ is a theorem of T. But then by (5).

(7) θ is a theorem of T.

Thus T is inconsistent. QED.

Thus CT^*, and so also the Conventionalist Thesis, are shown conclusively to be false. This returns us to our original question. Where does the mathematician get the raw material from which he begins his proofs if not from axioms viewed as stipulative conventions? To say that the axioms are the result of conventional stipulation is in effect to say that the mathematicians made them axioms. But this is not the case, as we have just seen. The axioms are not man-made. The axioms are true independently of anything that mathematicians may do. The mathematician does not just make them true: he somehow *comes to know* that they are true, and then uses them to show that other things are also true. But now, *how* does he come to know that these axioms are true? He doesn't prove them. They are what are left over after everything else has been proven. And he doesn't discover them inductively. The only thing that is left is that he simply *sees* that they are true, by some sort of intellectual vision or intuition. This is in fact just what mathematicians generally say that they are doing. This is what has traditionally been called "mathematical intuition." In proving mathematical theorems, the mathematician relies upon his mathematical intuitions and draws from that body of mathematical truths which he takes to be "self-evident," and these go beyond any stipulative axioms. This notion of self-evident mathematical truth is very unclear, but it is only by clarifying it that we can get clear on what a mathematical proof is.

Lest it be thought that this shows that mathematical provability transgresses logic and analytic truth, and hence that mathematics is somehow "synthetic *a priori*," let me hasten to add that theorems exactly analogous to the above can also be proven for set-theory and for higher-order logic. The situation is no better there. No axiomatization of these logical theories can exhaust the set of theorems we can prove classically within them. And as the provable theorems of logic (including higher-order logic) are surely the clearest cases of any interesting concept of analytic truth, this shows that analyticity itself cannot be viewed as the result of stipulative conventions. Just as the mathemtician must have recourse to mathematical intuitions, so in logic we must have recourse to "logical intuitions." These are in fact one and the same thing.

The thought that mathematical and logical proof are based ultimately on intuition seems to be repugnant to many philosophers. And yet, I think the conclusion is inescapable given the above argument. It will not do simply to maintain that intuitions could not possibly provide good grounds for a statement that something is a truth of mathematics, because in point of fact, this is just the way mathematicians do prove things.[9] But there is no reason to think that these mathematical intuitions are in any way mysterious. We must have some sort of mental faculty which allows us to intuit these things, and this mental faculty is presumably just as capable of physiological explanation as is our faculty of sight or of hearing. It may even be possible to give an account of our mathematical intuitions in terms of some other quite mundane mental process, such as imagination.

---

[7] This theorem can be obtained easily from Lemma 3 of Richard Montague's "Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability," *Acta Philosophica Fennica*, vol. 16 (1963), pp. 153–167.

[8] The construction is as follows: First we construct the diagonal function $\delta(x)$ such that if $f(z)$ is any formula of the language of T having the single free variable z, then $\delta(\sharp f(z)) = \sharp f(\sharp f(z))$. Then we construct the function $Neg(x)$ such that if ϕ is any formula of the language of T, $Neg(\sharp\phi) = \sharp{\sim}\phi$. Then let $k = \sharp\mathfrak{P}(Neg(\delta(z)))$, and $\theta = \mathfrak{P}(Neg(\delta(k)))$. Then $\sharp\theta = \delta(k)$ is a theorem of P, and so $\theta \equiv \mathfrak{P}(\sharp{\sim}\theta)$ is a theorem of P.

An objection that has frequently been raised to intuitionistic ethics will no doubt also be raised here. What happens when different people's intuitions disagree? How is it possible to decide who is right? But I think that these questions are quite easily answered. Our mathematical intuitions can profitably be compared with our faculty of sight. Mathematical intuitions don't provide us with incorrigible grounds for mathematical judgments, any more than our sight provides us with incorrigible grounds for judging the colors of things. If something looks red to us, then we have *prima facie*, but not conclusive, grounds for thinking that it is red. And similarly, if our mathematical intuitions tell us that something is true, we have *prima facie*, but not conclusive, grounds for thinking that it is true. We can perfectly well recognize a mathematical intuition to be incorrect, just as we can recognize that something is not the color that it looks to us. This is done by relying upon other mathematical intuitions and using them to prove that what we thought we intuited to be true was really false. A particularly clear case of this is the Axiom of Comprehension in set theory. Almost everyone's intuitions seem to be in agreement that given any set-theoretic predicate $\phi(x)$, there exists a set of all objects satisfying this predicate. But Bertrand Russell showed that this principle led to a contradiction, and thus was false. His proof was itself grounded on intuitions, so he was using intuitions against intuitions to show that certain intuitions were incorrect. In precisely the same manner, we use facts discovered by relying upon our sight to show that certain things are not the way they look to us. For example, we may judge that something that looks red to us is not really red but just looks that way because there are red lights shining on it, but in order to discover the connection between the apparent color of something and its illumination by colored lights we had to rely upon our sight elsewhere. Thus our mathematical intuitions, like our sight, provide us with *prima facie* grounds for mathematical judgments, but these *prima facie* grounds can be overridden in a particular case by using other mathematical intuitions *against* a particular intuition.[10]

The conclusion that mathematical proof is based ultimately on intuition has interesting implications in several areas. First, let us consider the controversy over whether mathematicians *discover* mathematical truth or *create* mathematical truth. This is often equated with the controversy between platonism and nominalism. The argument is that if mathematicians discover mathematical truth, then there must exist independently some mathematical entities, such as numbers, for these truths to be about. On the other hand, if mathematicians create mathematical truth, this must mean that mathematical truth is simply the result of arbitrary stipulation concerning what we are going to take to be true in a mathematical theory, and so a mathematical theory is just an uninterpreted formal system, with the mathematical terms devoid of any reference to independently existing mathematical entities.

Given the above understanding of what it would be for mathematicians to create mathematical truth, the failure of the Conventionalist Thesis also shows that mathematicians do not create mathematical truth. The mathematical and logical structure of the universe is there, and we discover it by intuition and by proofs based on intuition. We do not create it by stipulation.

I think our results also have some implications for Hilbert's program. Hilbert was worried that that part of mathematics which makes essential use of infinite totalities somehow has a less secure foundation than the strictly "finitary" parts of mathematics. The possibility arises that the adjunction of infinitary mathematics to finitary mathematics may allow us to prove with infinitary methods the denials of certain theorems we can prove with finitary methods, and thus lead to a contradiction. If this is *not* the case, then the adjunction can at least cause no troubles for finitary mathematics and it provides us with interesting new concepts to investigate. But if it is the case—if the adjunction does lead to a contradiction—then we must reject infinitary mathematics and content ourselves with strictly finitary mathematics. Hilbert proposed to prove, using finitary methods, that no such contradiction can ensue from the adjunction.

---

[9] Cf. N. Bourbaki, "The Architecture of Mathematics," *American Mathematical Monthly*, vol. 57 (1950), p. 227.

[10] This state of affairs can be described precisely by saying that our mathematical intuitions provide the *criterion* for our judgments that something is self-evidently true, in the technical sense of "criterion" defined in my "Criteria and Our Knowledge of the Material World," *The Philosophical Review*, vol. 75 (1966).

It is generally supposed that Gödel's second incompleteness theorem dealt the death blow for Hilbert's program, because it showed that to prove the consistency of any fairly powerful axiomatic theory, one must use a still more powerful theory. Thus, for example, one cannot prove the consistency of an axiomatic set theory within that axiomatic set theory. One must go to an essentially stronger theory. And I think that so long as we are thinking only of axiomatic theories and proving their consistency, as indeed Hilbert was, this argument is quite correct. However, one might well maintain that Hilbert was only talking about axiomatic theories because he supposed, incorrectly, that mathematics could be completely axiomatized. As this is now known to be false, it would seem reasonable to talk instead about proving the consistency of a "branch" of mathematics, such as set theory, rather than talking about proving the consistency of *axiomatic* set theory. And we have just seen that we can prove things within elementary number theory that we cannot prove within any axiomatic theory. Thus it is not unreasonable to suppose that we may be able to prove the consistency of, say set theory, within elementary number theory. Gödel's theorem does not preclude this possibility. And it would seem quite reasonable to call such a consistency proof "finitary." Thus I think that the above refutation of the Conventionalist Thesis can be taken as breathing new life into the Hilbert program.

Let us consider what this all means for the thesis that all *a priori* or analytic truth results from "rules of language." If all it means to say is that a statement which expresses a rule of language is analytic, then of course this thesis is true, but then it cannot be taken to *explain* analytic or *a priori* truth. On the other hand, if rules of language are supposed (as they generally are) to represent linguistic conventions, then the above argument shows that *a priori* truth cannot be explained in terms of rules of language. Language, as a human artifact, really has very little to do with *a priori* truth. Language *expresses* truth; it does not create it.

It seems, then, that in the light of the failure of the Conventionalist Thesis we must rethink a number of the generally accepted beliefs in the philosophy of mathematics and the philosophy of logic.

State University of New York at Buffalo

Received June 29, 1966

American Philosophical Quarterly
Volume 4, Number 3, July 1967

VI. RELIGOUS DISCOURSE AND RATIONAL PREFERENCE RANKINGS

J. W. SWANSON

IT is now some twenty years since John Wisdom first raised the question of non-experimental hypotheses in his hauntingly beautiful essay entitled "Gods."[1] In that essay he dissected the relation of fact to non-falsifiable theory—i.e., to the sort of theory against which one allows no new empirical findings to count. That a hypothesis has ceased to be experimental—in the sense that no fresh evidence is permitted to count as invalidating or disconfirming it—does not warrant the claim that the hypothesis "has nothing to do with the facts." To say this would be to distort what is ultimately an enormously complex relationship, Wisdom urges. Here one recalls that his arguments have as their focus a parable—one much too familiar to need detailed recounting here—concerning a pair of rival hypotheses, based on identical evidence, about the state of a garden and the possibility of its being cared for by a gardener. The purport of Wisdom's discussion concerning the rival hypotheses was that although the competing gardener and non-gardener theories were not experimental—i.e., no new evidence would be allowed as unfavorable by exponents of either theory—they nevertheless *made a difference*, indeed a very great difference. What I should like to do in this essay is to attempt to spell out, in more detail and from another perspective than Wisdom's, some of the ways in which non-experimental hypotheses, especially with regard to their use in religious discourse, do make a difference, do "count" so to speak, in any general rational construction of the way things are. The issue is an important one, and a neglected one. For although a number of first-rate philosophers have addressed themselves to this problem in the last twenty years, matters are still pretty much where Wisdom left them in 1945.[2]

Let us follow R. M. Hare[3] in referring to that respect in which one person differs from another in a non-disconfirmable way, as that person's "blik" with respect to the matter under disagreement. Let us further refer to unfalsifiable sentences used to make (putatively) non-analytic assertions about the general nature of things—sentences, incidentally, which are ordinarily classified as "metaphysical," "theological," or "metaphorical" in character—as "blik-induced sentences." Examples of blik-induced sentences are: "God loves us as a father loves his children" (Flew), "Every event has its cause," and "The universe and everything in it just doubled in size." Thus, a specific blik may inspire or induce a number of sentences, and these may collectively be said to constitute a metaphysics or a theology. The set of sentences that includes "God loves us as a father loves his children," "God gave his only son to save man," and "Mary gave birth to Jesus by immaculate conception," for example, may be said to be induced by a particular blik, and to constitute jointly a theology, the one usually associated with Christian belief.

Recall that Hare introduced the notion of a blik as a rejoinder to an argument of Flew[4] to the effect that putative theological assertions are not assertions at all, since nothing is allowed to count against them. It does not follow, Hare argued—and here his views are clearly allied to those of Wisdom—that just because certain statements (ones about gardeners, for example) are not assertions, that they do not count, do not make any difference. To make clear that they *do* count, Hare coined the expression "blik" to identify the respect in which two persons can disagree with one another over some crucial issue, while at the same time being in full agreement about the facts concerning it. And if bliks are not the sort of things that can be falsified, Hare urged, they none the less must be taken into account in any adequate analysis of non-formal argument. Now in a way, it is no more than common sense to say that bliks make a difference, that they do count. Paul's post-epiphanic conduct was sufficiently different from what preceded it to change the entire course of history, as Hare himself points out elsewhere.[5] *That* blik certainly counted. So in a certain sense it is entirely uncontroversial and unexceptional to claim, as Hare does, that bliks make a difference, irrespective of the non-falsifiability of the associated blik-induced sentences. What is astounding is that, given the obvious importance of the concept of the blik, professional philosophers—apart from the handful concerned with philosophical theology[6]—have paid so little attention to it. Clearly, what is sorely wanted at the present juncture is some sharpening of the informal logic of non-experimental hypotheses. In what follows, I shall attempt to supply some analytical tools requisite for this task, with particular regard to the role of religious discourse.

---

[1] Reprinted in *Logic and Language*, Antony Flew (ed.) (Oxford, Blackwell, 1951).

[2] Perhaps this is too strong. Excellent work on this problem has been done by both J. W. N. Watkins; cf. in particular his "Confirmable and Influential Metaphysics," *Mind*, vol. 68, (1958), pp. 344–365, and by R. N. Hanson, *Patterns of Discovery* (Cambridge, Cambridge University Press, 1958). Ultimately, of course, the inspiration for all current papers on non-experimental hypotheses harks back to Popper's seminal work on unfalsifiability. Cf. Karl Popper, *The Logic of Scientific Discovery* (New York, Basic Books, 1959), in particular ch. 1, sect. 6.

[3] R. M. Hare, §B of "Theology and Falsification" in *New Essays in Philosophical Theology*, Antony Flew (ed.), (London, SCM Press, 1955); and "Religion and Morals" in *Faith and Logic: Oxford Essays in Philosophical Theology*, Basil Mitchell (ed.) (Boston, Beacon Press, 1957).

[4] Antony Flew, §A of "Theology and Falsification" in *New Essays in Philosophical Theology*, *op. cit.*

II

Let us assume for the purposes of the logical model to follow that the persons x, y, z referred to are rational in their preferences. That is, we shall assume that every individual with whom we are concerned would exhibit, for every appropriate finite set of sentences K which he is able to understand, a rational preference ranking (*RPR*).[7] This is to say that an experimenter E could, by interrogation or observation (or a combination of both) of an individual x, determine with respect to two sentences a and b whether x preferred a to b, b to a, or was indifferent with respect to a and b. The exact characterization of the term "preference" or "prefers," but not its logical properties, could be expected to vary with the inquiry at hand. Generally speaking, one would expect to find in any axiomatization of that part of the theory of pragmatics[8] dealing with the logic of preference, some four place predicate "$Pxabt$" taken as a primitive expression of the formal theory and standing for some such pre-systematic locution as "x considers a better confirmed (verified, substantiated) than b at time t,"or "x prefers to base his actions on a as opposed to b at time t," or "x chooses a over b at time t," etc. For our purposes, however, the following negatively cast version will turn out to be the most useful: "x is less willing to relinquish a than b from his corpus of knowledge at time t." The reasons for this choice will come out in the sequel. Of course this definition will restrict any set of sentences K ordered by the preference relation to just those about which an individual purports to have some knowledge, ruling out such purely aesthetic choices as, say, color preferences. But this is harmless; after all, we are trying to capture the notion of a *rational* set of preferences, not just any set. For the moment, let us merely emphasize that the most important feature of the preference relation, apart from its logical properties described in the next paragraph, is that it must be possible for E to determine from x's *behavior alone*—including, of course, his verbal behavior—whether x is less willing to relinquish a than b at time t.

Thus the experimenter E would single out some candidate for blik analysis—e.g., "determinist," "Christian," "atheist"—and some class of sentences K presystematically associated with that candidate. By observing and interrogating the individual x, E would be able to establish x's *RPR* with respect to K. If x's *RPR* hews to the main lines of the commonplace and presystematic use of the term under analysis, then E could reasonably claim that the *RPR* is a precise and clear characterization of the attitude embedded in the use of the presystematic term. It should be clear from what has been said that the value of any experimental implementation of the notion of *RPR* with regard to a particular body of discourse will hinge largely on the prior selection of K, a selection the adequacy of which might be measured by the uniformity of the *RPR*'s that it determines for certain classes of individuals imputed to share some common belief.

---

[5] Hare, "Faith and Logic," *passim.*

[6] Their work has been generally of high caliber, full of insight.

[7] For a clear account of the notion of rational preference ranking, cf. Donald Davidson, J. C. C. McKinsey, and Patrick Suppes, "Outlines of a Formal Theory of Value I," *Philosophy of Science*, vol. 22 (1955), pp. 140–160; and R. M. Martin, *Intension and Decision* (Englewood Cliffs, N.J.; Prentice-Hall, 1963), pp. 45 ff.

[8] A formal theory of pragmatics has been developed by R. M. Martin in *Toward a Systematic Pragmatics* (Amsterdam, North-Holland Publishing Co., 1959).

Preference, of course, must be considered as transitive ("$Pxabt$ & $Pxbct \supset Pxact$" must hold in the axiomatization) and asymetrical ("$Pxabt \supset \sim Pxbat$" must hold). Indifference, which incidentally can be defined in terms of preference ("$Ixabt$" for "$\sim Pxabt$ & $\sim Pxbat$") also must be transitive, and symmetrical as well. Finally, to say that x exhibits a *RPR* for a class K of sentences at time t, exactly one of the following must obtain: x prefers a to b at t, x prefers b to a at t, or x is indifferent with respect to a and b at t. It is clear that the notion of *RPR* is essentially a formal characterization of the informal notion of grading. But the explicated notion of grading turns out to be a powerful tool of analysis,[9] as we shall see.

Let us make the further simplifying assumption for the model that the time t at which individual x ranks a and b may be of any finite duration. (Perhaps we would do better to read "$Pxabt$" as "x prefers a to b *during* t.") This allows us to treat x's choices as dispositions. Of course, in the real world, individuals change their minds with alarming and unsettling frequency. Still, most individuals, I suppose, are fairly constant—or would be under interrogation—with respect to their preferences with respect to Scotch and Bourbon, monogamy and polygamy, etc. The switch, incidentally, from talk about preference of sentences to talk about preference of things, concrete or otherwise, is justified by the assumption that preference in the latter corresponds in a fairly obvious way to preference in the former. Suppose x prefers monogamy to polygamy, in the sense that he prefers a state of affairs (world) in which monogamy is realized to one in which polygamy is realized. If confronted by E with "Monogamy is a useful institution" and "Polygamy is a useful institution," then in a *RPR* over some K containing both sentences, x would surely rank the former ahead of the latter. The point is that information about x's preferences among sentences furnishes excellent grounds for inferences about x's preferences concerning things, states of affairs, etc.[10] And *these* preferences can surely be treated as dispositions. Thus, if we let t be k units of time in length (k hours, days, etc.), then to treat a given *RPR*—for a specific individual x and a certain class of sentences K—as a disposition is to suppose that during the interval t, x's choices remain constant. Finally, I suggest that we simply identify bliks with dispositions as characterized by *RPR*'s. To have a certain blik is to have a disposition, an attitude if you prefer, determined by and associated with a certain *RPR*. This is the crux of the model I am proposing for the analysis of bliks. In what follows, I shall explore some of the consequences of looking at the old problems raised by Wisdom and Hare from the fresh vantage point of the model.

But first let me make it clear that I do *not* pretend that the preceding outline can serve as an adequate, in the sense of complete, account of religious belief (as generally expressed in religious discourse) *for the religious believer*. Rather it will be used in the next section as a partial and *neutral* account of the logic of religious discourse; partial, because certainly the religious believer will find it deficient insofar as it fails to account for some sort of prerogative cognitive status for particular religious or theological sentences (e.g., "God exists"), irrespective of how one ranks them; neutral, because both religious believer and sceptic alike can accept this account of linguistic behavior regarding discourse associated with religious belief. Indeed, in so far as the *RPR* account of religious discourse is merely commonplace knowledge regularized and made explicit, the most intensive "believer" can accept my model of religious belief as unthreatening, although he will undoubtedly not accept it as a *complete* account. After all, there is no extravagant claim to revelation in the above account. What is offered is a regimented version of dispositions or attitudes in terms of a model. But models, even logical ones, are always susceptible of being taken too seriously. To avoid these dangers, we must recall that the chief function, perhaps the only important function, of a model should always be the heuristic one of furnishing new *conceptual* insight, of allowing one to view the old from an intellectually unaccustomed vantage point, and in terms of novel explanatory principles. Let me turn in the concluding section, then, to what I believe to be the five most important insights suggested by the *RPR* model for religious discourse.

---

[9] It has good uses elsewhere, as well. Cf. my "On Models," forthcoming in *The British Journal for the Philosophy of Science*.

[10] Cf. Martin, *op. cit.*, p. 47.

III

First, and here I am only underlining the obvious, the *RPR* model of religious belief in terms of dispositions singles out a measurable, empirical, and quite probably useful scientific property of religious belief. A religious attitude, a blik, is thus as measurable as any sort of behavior is. And if to look at them as functions giving neat orderings of certain kinds of sentences is to some extent a distortion, it is, I submit, much less a distortion than to construe them as occult entities, beyond the reach of measurement and analysis. But is an interpretation of attitudes (bliks) in terms of a *RPR* model really a distortion at all? Hare's own position, as well as those of other analysts of religious discourse, can, I believe, be construed as actually implying something very much like the *RPR* interpretation. "The person who worships is bound to govern his conduct (or let it be governed) in a certain way; and he is also bound to believe in the truth of certain factual statements (empirical ones about what has actually happened in the world, and what is likely to happen)."[11] But surely, to "believe in the truth of certain factual statements" is to prefer these statements to others (for example, to their negations), to rank them ahead of a multitude of other statements, to be inclined to bet on them, base one's conduct on them as opposed to other statements, and so forth. Or take Ronald Hepburn's remark that "the parable and its associated pattern of behavior legislate not for any fraction of the believer's life, but for every aspect of it [commanding] . . . his total imaginative vision of the nature of man."[12] What is it for an attitude to "legislate" for "every aspect" of one's life if not to imply clear-cut choices in one's preferences among those sentences referring to the aspects of life which are of the greatest concern?

Secondly, to construe religious discourse in terms of the *RPR* model is to make it possible to incorporate the separately unfalsifiable sentences of religious discourse into a corpus in which they assume status and position relative to other sentences, not all of which need be religious in character. Religious sentences on this construction would no longer be expected to have meaning in isolation, but only in company both with their fellows and with sentences of other domains of rational interest. There is a distinct resemblance here between the *RPR* construction and the view of Quine that ". . . our statements about the external world face the tribunal of sense experience not individually, but only as a corporate body."[13] Just as Quine's doctrine shifts the focus of attention from theory-sentences to the theory itself, so the analysis of religious discourse in terms of dispositions to certain *RPRs* shifts the focus from isolated unfalsifiable sentences to the relation such sentences bear to all other sentences of the corpus constituting an individual's beliefs. Consequently, to see religious attitudes as dispositions toward certain *RPR*s is to free oneself from the fetish of the unfalsifiable. For first of all, *RPR*s are not themselves statements, hence are not the sorts of things that are qualifiable as true or false. So Hare, on our account, is surely correct in tagging bliks as non-falsifiable. But the weakness of positions such as Hare's has always been that they purchase unfalsifiability of religious utterances at the price of stripping religious discourse of all claim to cognitive, or even rational status. On Hare's view, to utter the blik-induced sentence "God loves us as a father" is not to make a statement at all, insofar as nothing would be allowed to count against it. Such an utterance is construed by Hare (at least as I read him) as no more than the exhibition or the evincing of an attitude, one that goes along with a certain theology. But is the almost universal belief on the part of users of religious language that the sentences they use *do* make statements—is this belief necessarily incompatible with Hare's account? Not, I think, if bliks are construed as *RPR*s. For viewed in that guise, surely our hypothetical experimenter could produce the sentence "This ordering is a reliable one" (one that ought to be followed) and "This ordering is not a reliable one." Then to say that one can determine from *x*'s *RPR* which of the two preceding sentences he prefers is to assimilate the heart of the cognitivist view to my model. The religious believer will, of course, prefer his own *RPR*, with respect to a class of sentences *K*, to that of the non-believer (perhaps this is even a tautology), and the assertion to that effect will itself be ranked in another *RPR* with respect to a class *K'* of sentences containing it and its negation, say, in addition to the members of *K*. And although a blik, *qua RPR*, is certainly not itself a hypothesis or theory, it does *rank* sentences which may be said to represent theories or hypotheses. So construed, the expression of religious belief comes to be not so much a matter of individual sentences as a question of interrelationships among a whole body of sentences.

---

[11] Hare, "Religion and Morals," *op. cit.*, p. 189.

[12] Ronald Hepburn, *Christianity and Paradox* (London, C. A. Watts and Co., 1958), p. 195.

[13] "Two Dogmas of Empiricism" in *From a Logical Point of View* (Cambridge, Mass.; Harvard University Press, 1953). p. 41.

Thirdly, construing religious belief in terms of attitudes makes sense of the many extremely diverse sorts of things that count in favor of (or against) religious belief. If we no longer think of religious belief in terms of isolated propositions, but rather as pervasive attitudes that exhibit themselves in terms of the importance that one assigns to the elements of his experience—an importance measurable, of course, by his disposition to rank sentences in a certain order—then the relevance to religious discourse of several different *kinds* of language use becomes apparent. The aesthetic and mythopoeic elements of religious belief, elements which inevitably appear negligible when one is concerned primarily with evidence for or against isolated propositions, become prominent. William Kennick has remarked on this important aspect of religious belief in the following words: "I suggest that the distinctive purpose of religious discourse is to articulate, arouse, sustain, and modify attitudes. . . . We may use ethical statements for this purpose, emotive statements, poetic statements, and so forth . . . religious discourse does use all of these forms of utterance as component elements."[14] If belief in God is a matter merely of believing the truth of certain assertions, then only such things as the cosmological argument and the argument from design could be persuasive in converting the unbeliever. But if religious belief is a function rather of attitude, then such things as Handel's "Messiah" are as likely to influence religious belief as learned journal articles.

Fourthly, Wisdom writes: " 'A difference as to the facts,' 'a discovery,' 'a revelation,' these phrases cover many things. . . . Things are revealed to us not only by the scientists with microscopes, but also by the poets, the prophets, and the painters. What is so isn't merely a mater of 'the facts'."[15] And later, after having introduced the parable of the men with their differing hypotheses about the garden, he asks: "What is the difference between them? The one says 'A gardener comes unseen and unheard. He is manifested only in his works with which we are all familiar,' the other says 'There is no gardener' and with this difference in what they say about the gardener goes a difference in how they feel toward the garden, in spite of the fact that neither expects anything of it which the other does not expect."[16] Wisdom's chief point, oversimplified a bit, is that the two men differ in spite of the fact that they agree with respect to all the facts about the garden. Suppose, now, that we look at their agreement about the facts and their ultimate difference of point of view in terms of the model. Let a_i be any sentence in the class K of n sentences $a_1, a_2, \ldots, a_n$ enumerating the facts, whatever they are, about the garden, and let $\sim a_i$ be its negation. Take all the $\frac{n!}{2!\,(n-2)!}$ pairs $\{a_i, \sim a_i\}$ and interrogate or observe x's and y's preference behavior with respect to these pairs. To say that x and y agree with respect to the facts at time t is to say that for all a_i "$Pxa_i \sim a_it$" is the case if and only if "$Pya_i \sim a_it$" is the case. Yet obviously x and y may still differ with respect to their *ranking* of any two distinct sentences a_i and a_j $(i \neq j)$. The difference in belief and attitude thus stands out with startling clarity in terms of the model.

Fifthly, there is another way in which bliks "count" and which comes out nicely in terms of the model. We can say that bliks are constitutive, in something like the Kantian sense, of the facts; that they determine in a very real sense just what it is that we are going to allow to count as facts. Now this position is not new. It is a recurrent leitmotif in the writings of almost all philosophers who have been influenced by the later work of Wittgenstein, in particular by his distinction between "seeing" and "seeing as." Indeed, Wisdom's penetrating parable of the garden is perhaps best understood as a development of this theme. It is central to Hare's own position: ". . . without principles of some sort we do not get any facts; there is no distinction between fact and illusion for a person who does not take up a certain attitude to the world."[16] This is the view that what we allow to count as a fact is a function of antecedently accumulated conceptual machinery—of a system of categories, explanatory principles, concepts, and labeling rubrics which somehow manage to sort, sift out, and arrange what we get from our nerve ends into a coherent conceptual pattern permitting of extrapolation to future experience. What I have in mind here is the sort of thing that Hanson captured so neatly when he said that "seeing is a 'theory-laden' understanding" and that "observation of x is shaped by prior knowledge of x."[17] Consider a simple case, the puzzle of the line drawing with the face hidden in it. Suppose one does not see the face initially, but upon prompting—i.e., upon being equipped with certain conceptual apparatus "face," "there in the tree," "knot-hole is one of the eyes," etc.—one does catch it. Then obviously, under interrogation, one's relative ordering of "there is a face in this picture" would shift with respect to other sentences about the picture.

---

[14] "The Language of Religion," *The Philosophical Review*, vol. 65 (1956), pp. 66–67.

[15] "Gods," *op. cit.*, pp. 192–193.

[16] "Religion and Morals," *op. cit.*, p. 190.

But so for genuine theories as well. To be told that light and micro-particles have wave properties is to be alerted for the appearance of diffraction bands, for evidence of a medium in which the waves can appear. Is it unlikely, then, that under the influence of a wave hypothesis about certain entities, the experimental physicist would look for certain kinds of evidence as opposed to others, and that this sort of preference would be reflected in some *RPR* containing such sentences as "There must be evidence of a medium for the waves"? *Mutatis mutandis*, these remarks carry over with equal force to religious language. For in so far as that sort of discourse purports to treat of that which is of ultimate concern, it must contain principles constitutive of what is to count as "ultimate," as of fundamental importance. And surely such overarching principles cannot fail but warp and realign our choices not only of facts in some naive sense as regards visions, miracles, etc., but more importantly, of our second-order preferences as well, of what is to *count* as evidence, as a fact, as "truth" even. And these preferences will be just as determinable, just as measurable in terms of *RPRs*, as are any of the others.

University of Massachusetts

Received June 6, 1966

---

[17] R. N. Hanson, *Patterns of Discovery* (Cambridge, Cambridge University Press, 1958), p. 19.

American Philosophical Quarterly
Volume 4, Number 3, July 1967

VII. THE CARTESIAN CIRCLE REVISITED

GEORGE NAKHNIKIAN

IN the Fourth set of Objections Arnauld wonders if Descartes' procedure in the *Meditations* is free of circularity: "The only remaining scruple I have is an uncertainty as to how a circular reasoning is to be avoided in saying: the only secure reason we have for believing that what we clearly and distinctly perceive is true, is the fact that God exists. But we can be sure that God exists, only because we clearly and distinctly perceive it; therefore prior to being certain that God exists, we should be certain that whatever we clearly and evidently perceive is true."[1]

The circle seems to be this:

(1) God exists
Therefore,
(2) Whatever I perceive clearly and distinctly is true.
But (3) I perceive clearly and distinctly that God exists.
Therefore,
(4) God exists.

I believe that a circle of this sort does exist in Descartes' *Meditations*, but it exists if, and only if, we assume that a certain aberrant view is part of Descartes' considered doctrine. The aberrant view is that even the most clearly and distinctly intuited propositions are not really known to be true except by those who know that God exists and that He cannot be a deceiver. This applies to *all* propositions except those that simply report or describe the clearly and distinctly perceived modifications of our own consciousness; it does not exempt the propositions, "I am conscious," "I exist," "If I am conscious then I exist," and "If I am in pain, then I am conscious." I can know that I am conscious only if I know that my being conscious is entailed by my being in pain, feeling heat, or any other immediately perceived mode of consciousness. But an evil demon may be deceiving me about the necessity of "If I am in pain, then I am conscious." Thus in the Third Meditation: "for as long as I am ignorant of this (namely that God exists and that He is not a deceiver), I do not see that I can ever be certain of anything else."[2] In the context of this utterance even the *cogito* propositions are not exempt. An omnipotent God, if he wants to, may "cause me to err even in those matters that I regard myself as intuiting . . . in the most evident manner (Latin)/ in a very evident manner (French)."[3]

There are places in Descartes' writings where he says that we can know certain propositions incorrigibly even if an evil demon is deceiving us about other things. For example, we can know incorrigibly that we, ourselves, exist. For how can the demon deceive *me* about anything at all unless *I* exist. So that we can know incorrigibly and without having to know that God exists such a conditional as this: If I am deceived at all, then I exist. However, in many passages, dating from 1630 to 1648, two years before his death, Descartes says or implies what amount to the aberrant view.[4] In a letter dated May 2, 1644, Descartes seems to be putting even something like the principle of noncontradiction under God's jurisdiction: God could have willed it "in general that it should not be true that contradictories cannot be together."[5] The overall impression is that Descartes is not wholly consistent. There is evidence that he is not altogether comfortable with the aberrant view, but there is also evidence that he never gave it up.

---

[1] AT (= Adam and Tannery, *Oeuvres de Descartes* (Paris, 1957)), vol. VII, p. 214 (Latin); vol. IX, p. 166 (French); HR (= Haldane and Ross, *The Philosophical Works of Descartes*, (New York, 1955)), vol. II, p. 92.

[2] HR, vol. I, p. 159. I have corrected the translation.

[3] AT, vol. VII, p. 36 (Latin); vol. IX, p. 28 (French); HR, vol. I, p. 158. The above translation is an improvement over HR. In "Descartes' Validation of Reason," *American Philosophical Quarterly*, vol. 2 (1965), H. G. Frankfurt is perfectly correct in finding the aberrant view in the *Meditations*. He is mistaken, I think, in not recognizing that it is aberrant. He is also mistaken in thinking that even with the aberrant view as a part of it, Descartes' doctrine can be defended against the Arnauld-type charge of circularity. I shall argue in this paper that that type of circularity arises if, and only if, the aberrant view is part of the doctrine.

[4] See Leonard G. Miller, "Descartes, Mathematics and God," *The Philosophical Review*, vol. 66 (1957), pp. 451–465, n. 3–8.

[5] AT, vol. IV, p. 118; Anscombe and Geach, *Descartes, Philosophical Writings* (London, 1954), p. 291).

The aberrant view is disastrous for two reasons: it generates a vicious circle, and it is intrinsically absurd by way of being self-stultifying.

Descartes is committed to the proposition that he must prove the principle of clarity and distinctness (the principle that all clear and distinct perceptions are true). He is further committed to the proposition that the principle can be proved only by being deduced from the proposition that God, who is not a deceiver, exists. Thus, in the Synopsis of the *Meditations* he writes: "... it is requisite that we know (Latin, *scire*; French, *savoir*) that all the things which we conceive clearly and distinctly are true in the very way in which we conceive them; and this could not be proved previously to the Fourth Meditation."[6]

I have corrected the HR translation by replacing "may be assured" with "know" in the sense of *scire* or *savoir*. This is important. Descartes uses "scire" as a technical term for knowledge of *theorems*, i.e., of demonstrated conclusions. In this application of the word, *scientia* is contrasted with *notitia* or *cognoscentia* (immediate apprehension of "first principles"). Descartes also uses "scientia" another way in which *scientia* is the opposite of *persuasio*. "When any reason whatever exists over and above what reasons we already possess that could impel us to doubt, we have mere *persuasio*. *Scientia* is persuasion from reasons so powerful that no reasons more powerful can shake them. And no one who is ignorant of God possesses such reasons as these."[7] The distinction between *scientia* and *persuasio* is a key to understanding the following important passage in the Reply to Objections II: "That *an atheist can know* (Latin: *cognoscere*, French: *connaitre*) *that the three angles of a triangle are equal to two right angles*, I do not deny, I merely affirm that, on the other hand, such knowledge (Latin: *cognitio*, French: *connaissance*) on his part cannot constitute true science (Latin: *scientia* [vs. *persuasio*], French: *science*) because no knowledge (Latin: *cognitio*, French: *connaissance*) that can be rendered doubtful should be called science (*scientia* vs. *persuasio*). Since he is, as supposed, an Atheist, he cannot be sure that he is not deceived in the things that seem most evident to him, as has been sufficiently shown: and though perchance the doubt does not occur to him, nevertheless it may come up, if he examines the matter, or if another suggests it; he can never be safe from it unless he first recognizes (*agnoscat*)" God.[8] The Latin and French insertions are mine.

In the light of this passage, in the lines quoted from the Synopsis Descartes is presumably saying that we must know (*scientia* vs. *persuasio* sense) as a theorem (*scientia* vs. *notitia* or *cognoscentia* sense) that the principle of clarity and distinctness is true. This suggests that even the "first principles" known by *notitia* or *cognoscentia* are no more than mere *persuasio* in the absence of knowledge (*scientia*, as the opposite of both *notitia* and *persuasio*) that God exists and is not a deceiver. And in saying that he cannot prove the principle of clarity and distinctness prior to the Fourth Meditation, Descartes is implying that the proof proceeds from the premiss that God exists. For the first argument for God's existence occurs in the Third Meditation. In the Fourth Meditation there is an argument purporting to deduce the principle of clarity and distinctness from God's existence. Thus, Descartes is undeniably wedded to the move from (1) God exists to (2) Whatever I perceive clearly and distinctly is true.

But this gets Descartes into a circle. For what is he saying? He is saying that he must know that (1) God exists, in order to prove that (2), the principle of clarity and distinctness, is true. But to know both as a theorem and as a matter of unshakeable conviction that (1) is true, he has to deduce it from premisses (axioms) which he knows as a matter of unshakeable conviction to be true. But, according to the aberrant view, no matter how clearly and distinctly he perceives the premisses for God's existence, he does not know that they are true unless he knows that the principle of clarity and distinctness is true. But to know that the principle of clarity and distinctness is true, he must know that God exists. We are in a vicious circle. We must prove that God exists, but in order to prove that God exists, we must assume that God exists.

Although Arnauld does not say enough for us to be sure, the circularity I have just described may have been the one he meant to single out.

In his reply to Arnauld, Descartes may seem to be tacitly disowning the aberrant view: "to prove that I have not argued in a circle in saying

---

[6] AT, vol. VII, p. 13 (Latin); vol. IX, p. 10 (French); HR, vol. I, p. 140.

[7] In a letter of May 24, 1640. AT, vol. III, p. 65, 1. 4–7.

[8] HR, vol. II, p. 39. AT, vol. VII, p. 141, 1. 3–13 (Latin). The Latin and French insertions are mine.

that the only secure reason we have for believing that what we clearly and distinctly perceive is true, is the fact that God exists; but that clearly we can be sure that God exists only because we perceive it, I may cite the explanation that I have already given at sufficient length in my reply to the second set of Objections, numbers 3 and 4. There I distinguished those matters that in actual truth we clearly perceive from those we remember to have formerly clearly perceived. For first we are sure that God exists because we have attended to the proofs that established this fact: but afterward it is enough for us to remember that we have perceived something clearly, in order to be sure that it is true; but this would not suffice, unless we knew that God existed and that he did not deceive us."[9]

In this reply Descartes refers to his replies to the Second set of Objections, numbers 3 and 4. The point under 4 has to do with showing that God cannot be a deceiver. That point is not immediately relevant. But the answer to 3 is relevant. This is what Descartes says: ". . . when I said that we could know (*scire*) nothing with certainty unless we were first aware that God existed, I announced in express terms that I referred only to knowledge (*scientia*) of such conclusions as can recur in memory without attending further to the proofs which led me to make them."[10] Here Descartes must be referring to the passages toward the end of the Fifth Meditation[11] where indeed he makes it clear that he is talking about conclusions that recur in memory. But this ignores the passages early in the Third Meditation where Descartes exempts nothing, not even the *cogito* propositions, from needing God's guarantee. By ignoring these passages, Descartes may create the impression that he is tacitly disowning the aberrant doctrine expressed in them. Indeed, as he now makes it clear, he distinguishes perception or apprehension (Lat: *notitia*, Fr: *connaissance*) of first principles, and knowledge (Lat: *scientia*, Fr: *science*) derived from deduction. ". . . apprehension (*notitia*) of first principles is not usually called knowledge (*scientia*) by dialecticians. . . ."[12]

Descartes goes on in this passage to say that the *cogito* propositions are apprehended intuitively. These propositions are "first principles." Our cognition of their truth is immediate, non-demonstrative. But this in itself does not support the supposition that Descartes is here disowning the aberrant view. For in the immediately following paragraph that we have already cited Descartes admits that an atheist can cognize that a triangle's internal angles all equal to two right angles, but without a knowledge of God, says Descartes, this cognition is mere *persuasio* and not true knowledge.

Descartes ought to have disowned the aberrant view not only to avoid circularity in his own system but also to avoid utter absurdity. According to the aberrant view, as long as Descartes is ignorant that God exists, he "doubts" all mathematical propositions. But to doubt *all* mathematical propositions is to doubt deduction itself.

This last remark needs explaining. There are two problems involving deduction, only one of which is relevant in the present context. One problem has to do with the fact that in a long deduction I have to *remember* at the last step that every previous step was intuitively certain. Related to this, there is, for Descartes, the problem that sometimes I *remember* having proved a theorem whose premisses or whose proof I do not remember at all. Descartes gives the impression that in his view these problems call for an argument which proves that clear and distinct memory impressions are true. In other words, these are fundamentally problems about the reliability of memory. That problem is not the one I am thinking of when I say that to doubt *all* mathematical propositions is to doubt deduction itself.

The fact about deduction which is relevant to this issue is that for every valid deductive argument there is a corresponding conditional which is necessarily true. In the aberrant passages Descartes says that without a knowledge of God even such simple and evident propositions as "All squares have four sides" may be reasonably doubted. It makes no difference that "we regard ourselves as intuiting them in the most evident manner."

But take "All squares have four sides." That proposition is equivalent to: "If anything is a square, then it has four sides." This is the conditional corresponding to the valid deduction-schema, "× is a square; therefore, × has four sides." Anyone who thinks that without a knowledge of God he can reasonably doubt that a square has four sides must, if he is to be consistent, concede that without a knowledge of God it is reasonable to doubt the deduction-schema. Hence, anyone who thinks that without a knowledge of God it is reasonable to doubt all mathematical propositions must also think that it is reasonable to doubt *all* deductions. To doubt that p entails q is not to know what to think about it. A man in that state of mind cannot tell a valid deduction from one that is not valid.

---

[9] AT, vol. VII, pp. 245–246 (Latin) vol. IX, Quatrièmes Réponses; pp. 189–190 (French); HR, vol. II, p. 114–115.

[10] AT, vol. VII, p. 140 (Latin); vol. IX, Secondes Réponses, p. 110 (French); HR, vol. II, p. 38. HR translate "de scientia earum conclusionum" as "the science apprehending such conclusions as." This is not strictly accurate. I prefer "knowledge of such conclusions as."

[11] HR, vol. I, p. 183–184.

[12] *Loc. cit.*

This generates a vicious circle in Descartes' system somewhat different from the one we have already described. It also leads to intellectual hara-kiri. The vicious circle arises as follows: Descartes wants to say that because he has *deduced* God's existence from certain premisses, he is "metaphysically"[13] certain that God exists. But how can he be "metaphysically" certain that his deductions are valid when, without God and God's goodness, not even necessarily true conditionals are "metaphysically" certain? In other words, Descartes *uses* deduction to prove that he knows with "metaphysical" certainty that God exists. But he wants to prove that he knows with "metaphysical" certainty that God exists, in order to prove, among other things, that he knows with "metaphysical" certainty that valid deductive inferences are "metaphysically" certain. This is viciously circular.

In the article cited in footnote 3 above, Frankfurt writes:

> It is evident that Descartes' argument does not suffer from the circularity with which it is commonly charged. Metaphysical doubt concerns the truth of what is intuited, and the removal of this doubt is effected without assuming that what is intuited is true. It is removed simply by the knowledge that a certain demonstration[14] has been successfully accomplished. This knowledge is, of course, the knowledge that certain things have been clearly and distinctly perceived. But it is not required that the *truth* of these things be supposed, and so the question is not begged.[15]

Unless I misunderstand this, it seems to be patently wrong. A demonstration exists only if (i) there exists a relevant valid deductive argument, and (ii) the premisses of this argument are true. Hence a man knows with metaphysical certainty that a demonstration exists only if he knows with metaphysical certainty that these two conditions are satisfied. But the man who holds the aberrant view cannot claim to know with metaphysical certainty either that a certain argument is deductively valid or that its premisses are true. It follows that such a man has no grounds for saying that he knows with metaphysical certainty "that a certain demonstration has been successfully accomplished." This is so because he has no grounds for claiming to know with metaphysical certainty that *any* demonstration has been accomplished. In other words, according to the aberrant view a man cannot know with "metaphysical" certainty that a certain argument is demonstrative unless he knows that God exists. But, according to Descartes, in order to know that God exists, a man must know with "metaphysical" certainty that a certain argument is demonstrative. I submit that Frankfurt's interpretation does not explain how Descartes can escape circularity if he commits himself to the aberrant view.

What is more, the aberrant view is tantamount to misologism. I repeat that any man who holds it has no basis for distinguishing between valid and invalid inferences, and a man who has no basis for thinking that he is making this distinction has no basis for thinking that he is reasoning. "Metaphysical" doubt about such analytic propositions as "All squares have four sides" is part and parcel of the aberrant view. But such doubts are radically self-stultifying. A man cannot philosophize without reasoning, and he cannot reason if he doubts even the simplest analytic propositions.

I have been arguing that the Cartesian Circle is generated *if* the aberrant view is assumed to be an integral part of Descartes' doctrine. I shall now explain why I believe that a circle is generated *only if* the aberrant view is assumed. Descartes wants to prove that God exists so that he can prove the principle of clarity and distinctness. Now suppose that we do not assume the aberrant view. That is, we suppose that whether or not we know that God exists, *some* clearly and distinctly perceived propositions are, and are known to be, true. God is needed so that we can deduce that *all* clearly and distinctly perceived propositions are true. Then it is open to Descartes to say that the premisses from which he deduces God's existence and veracity are clear and distinct propositions that are known to be true *per se*. A knowledge of God's existence and veracity is required, Descartes might say, to show that the following are true: all clearly and distinctly perceived perceptual propositions, all clear and distinct memory impressions, all clear and distinct perceptions of our own states of wakefulness, and all clearly and distinctly perceived mathematical propositions which are not self-evident to us.

---

[13] "Metaphysical" certainty is certainty that precludes all *reasonable* doubt. To know with "metaphysical" certainty is to have *scientia* vs. *persuasio*.

[14] The demonstration in question is "a proof that there is an omnipotent deity who is not a deceiver and whose existence, accordingly, entails that reason is reliable" (Frankfurt, *op. cit.*, p. 155).

[15] *Ibid.*

There are good reasons why we should not be satisfied with Descartes' attempts to prove these propositions. The proofs of God's existence, the proof of the principle of clarity and distinctness, and the attempt to explain the clarity and distinctness of memory impressions and perceptual propositions in terms of coherence criteria are riddled with problems, but circularity need not be one of them.[16] The system I find in the *Meditations* becomes circular if, and only if, we add to it the aberrant view that *no* clear and distinct intuition is known to be true except by those who know that God exists and that He is not a deceiver.

Wayne State University

Received May 2, 1966

---

[16] These problems are dealt with in some detail in Parts II and III of my book *An Introduction to Philosophy* (New York, 1967).

AMERICAN PHILOSOPHICAL QUARTERLY

Edited by

NICHOLAS RESCHER

With the advice and assistance of the Board of Editorial Consultants:

William Alston
Alan R. Anderson
Kurt Baier
Lewis W. Beck
Nuel D. Belnap, Jr.
Roderick M. Chisholm
L. Jonathan Cohen
James Collins
James M. Edie
Peter Thomas Geach
Adolf Grünbaum
Carl G. Hempel
John Hospers
Raymond Klibansky
Ernan McMullin
Benson Mates
John A. Passmore
Günther Patzig
Richard H. Popkin
Wesley C. Salmon
George A. Schrader
Wilfrid Sellars
Alexander Sesonske
J. J. C. Smart
Manley H. Thompson, Jr.
James F. Thomson
G. H. von Wright
John W. Yolton

VOLUME 4/NUMBER 4 OCTOBER 1967

CONTENTS

I. S. MORRIS ENGEL: *Wittgenstein's "Foundations" and Its Reception* . 257

II. R. M. MARTIN: *Facts: What They Are and What They Are Not.* . . 269

III. WILLIAM W. ROZEBOOM: *Why I Know So Much More Than You Do* 281

IV. CLIVE BECK: *Utterances Which Incorporate a Value Statement* . . . 291

V. HOWARD SMOKLER: *The Equivalence Condition* 300

VI. ALASTAIR MCKINNON: *"Miracle" and "Paradox"* 308

VII. DAVID P. GAUTHIER: *Moore's Naturalistic Fallacy* . . . 315

PUBLISHED BY BASIL BLACKWELL WITH THE COOPERATION OF THE UNIVERSITY OF PITTSBURGH

AMERICAN PHILOSOPHICAL QUARTERLY

POLICY

The *American Philosophical Quarterly* welcomes articles by philosophers of any country on any aspect of philosophy, substantive or historical. However, only self-sufficient articles will be published, and not news items, book reviews, critical notices, or "discussion notes."

MANUSCRIPTS

Contributions may be as short as 2,000 words or as long as 25,000. All manuscripts should be typewritten with wide margins, and at least double spacing between the lines. Footnotes should be used sparingly and should be numbered consecutively. They should also be typed with wide margins and double spacing. The original copy, not a carbon, should be submitted; authors should always retain at least one copy of their articles.

COMMUNICATIONS

Articles for publication, and all other editorial communications and enquiries, should be addressed to: The Editor, *American Philosophical Quarterly*, Department of Philosophy, University of Pittsburgh, Pittsburgh, Pennsylvania 15213.

OFFPRINTS

Authors will receive gratis two copies of the issue containing their contribution. Offprints can be purchased at low cost through arrangements made when checking proof.

SUBSCRIPTIONS

The price *per annum* is six dollars for individual subscribers and ten dollars for institutions. Checks and money orders should be made payable to the *American Philosophical Quarterly*. Back issues are sold at the rate of two dollars to individuals, and three dollars to institutions. All correspondence regarding subscriptions and back orders should be addressed directly to the publisher (Basil Blackwell, Broad Street, Oxford, England).

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 4, October 1967

I. WITTGENSTEIN'S "FOUNDATIONS" AND ITS RECEPTION

S. MORRIS ENGEL

I

NO one who has been following the reviews of Wittgenstein's posthumously published *Remarks on the Foundations of Mathematics* will have failed to be struck by the almost unanimous sense of disappointment and disapproval expressed by their writers. This unanimity is itself a remarkable phenomenon. It raises, however, a number of questions, foremost among them, whether perhaps the sense of the book has not somehow been mistaken.

It would be tedious and an imposition on the reader to try to discuss here all or nearly all of the reviews which have appeared to date. Fortunately, only quite recently three of the major ones (together with long selections from the *Foundations* itself) have been collected in a volume designed as a textbook for courses in the philosophy of mathematics[1] and I shall use it as my point of reference. Since, however, to select is in a way to interpret, it will not be possible to avoid commenting also on the choice of selections and the questions which such choices too tend to raise.

Instead of trying to attack the selections and reviews on several different fronts or issues, I shall perhaps be able to make my point best if I concentrate my attention on merely one. The issue I have chosen has to do neither with mathematics itself, nor even with what we might properly call "the philosophy of mathematics." It might perhaps best be described as dealing with the *psychology* of mathematics. This is not to say that it is therefore unrelated to philosophy, for running throughout all Wittgenstein's works are reflections which, although not in themselves philosophical, have, however, intimate and serious implications for philosophy. This is especially true, I believe, of his contributions here. Certainly they are not contributions to mathematics as such. He himself is careful to point this out. Whether they should be regarded as contributions to philosophy or psychology, and what is to be gained from distinguishing them in this way, is what I wish to explore here. It is the failure to recognize and acknowledge this distinction which, it seems to me, lies behind the unrepresentative character of the selections and the essentially unilluminating nature of the reviews which accompany them.

II

Before turning directly to these questions there are one or two other points which students of Wittgenstein are likely to find disturbing. For some years now many editors and prospective editors have been unable to persuade Wittgenstein's literary executors to permit them to reproduce selections from his works. Partly, the difficulty has been in complying with Wittgenstein's wish that any translation always be accompanied by the original; and partly in accommodating his desire to keep the text intact. (For with a philosophy which, at least in appearance, seems so new and unfamiliar, who can be certain that he has achieved such a familiarity with it as confidently to be able to take such liberties with it.) The present selections thus represent a departure on two counts, for not only are they excerpts from a larger work but are here given only in translation. Although this will no doubt be cheered and welcomed by all prospective editors, others will regard it as a disturbing development.

Lest this may seem rather harsh on the present editors, it should be added that these remarks apply with equal force to the work from which these excerpts are taken. That work is itself a collection of selections from a larger body of jottings. How well the editors did their job not even they can, of course, know and will not be known until such time as the whole work is made available to the public, as it inevitably will be. Unfortunately this is unlikely to happen in the lifetime of present readers. That is why it is unfortunate that the editors decided to offer the public a selection and not the entire mass of material. In such matters it is always better to have either nothing or everything. As things stand, what is available can only detract from Wittgenstein's stature, for the reader may be led to apply inappropriate criteria to it, being misled by the finished quality which a selection inevitably tends to give a work. In its original state no such impression is created to give rise to false expectations and feelings of disappointment.

---

[1] *Philosophy of Mathematics: Selected Readings*, edited with an Introduction by Paul Benacerraf and Hilary Putnam (Englewood Cliffs, New Jersey; Prentice-Hall, Inc., 1964). The three reviews reprinted in this text are, in order of their appearance in the text: Alan Ross Anderson "Mathematics and the 'Language Game'," *The Review of Metaphysics*, vol. 11 (1958), pp. 446–458; Michael Dummett, "Wittgenstein's Philosophy of Mathematics," *The Philosophical Review*, vol. 68 (1959), pp. 324–348; Paul Bernays, "Comments on Ludwig Wittgenstein's *Remarks on the Foundations of Mathematics*," *Ratio*, vol. 2 (1959), pp. 1–22.

I do not wish to suggest that the three reviewers whose papers are here reprinted were so misled. That they are disappointed with the work is obvious. "It is hard to avoid the conclusion," Alan Ross Anderson writes in his review, "that Wittgenstein failed to understand clearly the problems with which workers in the foundations have been concerned."[2] A page later he writes: "It is very doubtful whether this application of his method to questions in the foundations of mathematics will contribute substantially to his reputation as a philosopher."[3] Michael Dummett is careful to point out in his opening paragraphs that the present work "has to be treated as what it is—a selection from the jottings of a great philosopher."[4] He notes that the striking contrast between it and the *Philosophical Investigations* (which "expresses with great clarity many forceful, profound, and quite definite ideas"[5]) "is due entirely to the different origins of the two books."[6] Dummett cannot, however, also resist saying of some of Wittgenstein's remarks that they are "plainly silly,"[7] "extremely hard to swallow,"[8] "extraordinary difficult to take . . . seriously,"[9] "thin and unconvincing,"[10] etc. And finally, Paul Bernays is led to say (almost contemptuously) that "Wittgenstein argues as though mathematics existed almost solely for the purpose of housekeeping."[11] Or again: "Wittgenstein's method aims at eliminating a very great deal. He sees himself in the part of the free-thinker combating superstition. The latter's goal, however, is freedom of the mind; whereas it is this very mind which Wittgenstein in many ways restricts, through a mental asceticism for the benefit of an irrationality whose goal is quite undetermined."[12]

In the light of such evaluations one may wonder what led the editors to devote almost a fourth of their text to Wittgenstein. Perhaps anticipating this uneasiness on the part of the reader, they say that what led them to do this was not because "Wittgenstein is more important in the philosophy of mathematics, or even, necessarily, as important as such seminal thinkers as Brouwer, Hilbert, or Frege. Rather it is in view of the great influence that Wittgenstein had on Russell and on the positivists, the great influence that he is currently having on 'analytic philosophers,' and the great difficulty of interpreting his often cryptic remarks on the subjects with which we are concerned" that led them to do so.[13] Concerning the selection of the reviews, they state: "Our purpose in including the reviews of his book is to facilitate discussion of Wittgenstein's later ideas in the philosophy of mathematics by providing possible starting points; we certainly do not wish to give the impression that we regard these criticisms as definitive."[14]

Now although this may be all very well, it is hardly a very forceful justification for including either the selections or the reviews. It might have been very well had the editors chosen reviews of a different sort, but since the present reviews tend to throw the value of the excerpts into doubt, what justification remains for including them in the first place? With such "starting points," in other words, the reader may not think it worthwhile to go on. There remains, of course, Wittgenstein's "great influence" on positivists, analytic philosophers, etc., but that is hardly a proper justification for devoting over a hundred pages to him in a text attempting to deal with the philosophy of mathematics in a systematic way.

Is there any justification at all then for attaching this kind of importance to this last work of Wittgenstein? Are there perhaps certain themes in the work which have somehow escaped notice but which might make a profound difference to our understanding of it? I believe there are. These are, however, as we shall see in a moment, more in the nature of certain psychological rather than philosophical observations. Perhaps that is why they have generally been overlooked.

---

[2] *The Review of Metaphysics*, *op. cit.*, p. 457.
[3] *Ibid.*, p. 458.
[4] *The Philosophical Review*, *op. cit.*, p. 324.
[5] *Ibid.*
[6] *Ibid.*
[7] *Ibid.*, p. 330.
[8] *Ibid.*, p. 332.
[9] *Ibid.*, p. 333.
[10] *Ibid.*
[11] *Ratio*, *op. cit.*, p. 14.
[12] *Ibid.*, p. 6.
[13] *Ibid.*, p. 25.
[14] *Ibid.*

III

If we are to make room for them we have now to work, I believe, with a threefold division of the subject. There are, first of all, the observations or contributions of a strictly mathematical sort. There are obviously many such remarks in Wittgenstein's book. But we have, of course, Wittgenstein's own word that this is not his main interest. On the contrary, what he seems ostensibly to be doing is the *philosophy* of mathematics. This is a second-order philosophical discipline and just as legitimate as such other, related second-order philosophical investigations as the *philosophy of art*, the *philosophy of science*, the *philosophy of morals*, etc. Such disciplines are not designed to improve one's *knowledge* of art, or science, or morals (and much less so, of course, to make one more moral, or a better artist, etc.) but rather, by making explicit the foundations, goals, limits, and presuppositions of such studies, are designed to improve one's *understanding* of them. In view of this almost obvious and natural distinction, it is difficult to understand Michael Dummett's criticism that "there is no ground for Wittgenstein's segregation of philosophy from mathematics but that this springs only from a general tendency of his to regard discourse as split up into a number of distinct islands with no communication between them (statements of natural science, of philosophy, of mathematics, of religion)."[15] Furthermore, Wittgenstein himself, he says, occasionally "displays no timidity about 'interfering with the mathematicians'."[16]

These remarks of Dummett's are prefaced by the observation that Wittgenstein's contribution consists in mediating between rival schools of mathematicians regarding such things as their different conceptions of proof, their formulation, consequences, etc. "It seems natural to suppose," he therefore argues, "that the philosophical task and the mathematical go hand in hand, for the precise formulation of a conception is not irrelevant to deciding on its correctness, and unexpected consequences of adopting it may lead one to revise one's opinion as to its value."[17] But we might reply on Wittgenstein's behalf that, of course, this is only "natural" from the mathematician's point of view —and in certain areas not even from his. For he may indeed learn a great deal from the philosopher of mathematics that may be of immediate interest to him. But the philosopher of mathematics has a much wider task; for there are, after all, further matters to explore. (For example, the point—touched on in passing by Dummett himself—that although certain mathematical procedures are not perhaps *wrong*, they may still not possess the *interest* we are inclined to attach to them.[18]) It is not certain Dummett has grasped all the implications of this point. Certainly, as we shall see in a moment, Wittgenstein is doubtlessly right on such points, and Dummett is wrong in insisting that "philosophy and mathematics have nothing to say to one another" and that "no mathematical discovery can have any bearing on the philosophy of mathematics."[19] What is perhaps remarkable is that towards the end of his review Dummett himself seems to recognize that such matters may indeed "have nothing to do with mathematics in particular."[20] But such points, as we shall see in a moment, have to do more with the psychology of mathematics than with its philosophy.

Now if we are right in distinguishing mathematics from the philosophy of mathematics by saying that the latter is a second-order discipline devoted to improving our understanding of what mathematics ideally is or ought to be, then we can say that whereas the philosophy of mathematics is a normative study, the psychology of mathematics is a purely descriptive, empirical discipline dealing not with what mathematics ideally is, but with what our ordinary common understanding takes it to be and the consequences which issue from so understanding it. The distinction here is not unlike that generally drawn between logic and psychology. Although both, as we now tend to regard them, may be said to be studies of the "laws of thought," logic we regard as being concerned with the laws governing the relations among concepts or propositions, psychology with the laws governing mental processes of thinking (the one is regarded by us as being a descriptive discipline devoted to studying how thinking actually takes place, the other as a normative or prescriptive discipline devoted to studying the standards to which our thinking ought to conform).

---

[15] *The Philosophical Review*, *op. cit.*, p. 326.

[16] *Ibid.*

[17] *Ibid.*

[18] *Ibid.*

[19] *Ibid.*

[20] *Ibid.*, p. 348.

Assuming this distinction is appropriate, the question is: Is there much of any importance that can be said about the ways in which mathematics is regarded by our ordinary common understanding? If one approaches Wittgenstein's writings from this point of view he will be surprised how much space and thought he devoted to this topic. From this point of view both the present selections and the book from which they are taken are misleading—the selections because they tend to bypass most of Wittgenstein's comments on this matter, the book because it leaves out a good deal that is necessary and which can be gotten (at the moment) only from other sources. Although I have mainly in mind here such further sources of information regarding Wittgenstein's thought as notes taken down at this time by other students (Norman Malcolm's *Math Notes*, for example),[21] I have in mind also such further works as the *Philosophical Investigations*. (Many portions of this work originate from the same period of reflection.)[22] Whether the mass of material from which the *Foundations* was edited contains these further and necessary reflections, only the editors can tell us. They do inform us in their Preface that an Appendix dealing with surprise in mathematics was omitted. As we shall see in a moment, this would certainly be relevant.

IV

One of the themes which is almost indispensable for a proper understanding of Wittgenstein's remarks here has to do with his view regarding the way in which our minds are held captive by "pictures." A certain form of this theme goes back as far as his early work, the *Tractatus*, in which, as everyone knows, a picture-theory figures prominently. The interesting thing here is that although it is true that Wittgenstein gave up *that* picture-theory later on, he never really gave up trying to develop *a* picture theory or explore it in other directions. This is one of those threads in Wittgenstein's thought which has a tendency to break off at certain points but which somehow never completely vanishes from sight.

Now what is it for the mind to be misled or held captive by pictures? A rather striking example of what Wittgenstein had in mind here is to be found in Norman Malcolm's *Memoir* of Wittgenstein. Malcolm recounts how "at one of the at-homes, Wittgenstein related a riddle for the purpose of throwing some light on the nature of philosophy. It went as follows: Suppose that a cord was stretched tightly around the earth at the equator. Now suppose that a piece one yard long was added to the cord. If the cord was kept taut and circular in form, how much above the surface of the earth would it be? Without stopping to work it out, everyone present was inclined to say that the distance of the cord from the surface of the earth would be so *minute* that it would be imperceptible. But this is wrong. The actual distance would be nearly six inches. Wittgenstein declared that this is the *kind* of mistake that occurs in philosophy. It consists in being *misled by a picture*. In the riddle the picture that misleads us is the comparison of the length of the additional piece with the length of the whole cord. The picture itself is correct enough: for a piece one yard long would be an insignificant fraction of the length of the whole cord. But we are misled by it to draw a wrong conclusion. A similar thing happens in philosophy: we are constantly deceived by mental pictures which are in themselves correct."[23]

I should like to suggest now that when Wittgenstein came to look at the sort of things which people or philosophers were saying about mathematics it struck him that the same thing was true of them too.

In one of his lectures[24] he points out that although certain mathematical expressions lend themselves to misunderstanding, it is not the mathematicians who are misled by them. It is different, however, with the philosopher of mathematics, or the man in the street. Their interest in these matters initially arises not from anything internal or directly relevant to these investigations, but rather from the associations these expressions tend to arouse in their minds. What fascinates them are the pictures which the mathematician's talk about his work tends to conjure up. And these pictures are misleading. For one thing, they tend to make his work seem more important and more glamorous that it would otherwise seem to be. Moreover, although this may not be a bad thing (for without such glamor no one might have become interested in these problems to begin with), such pictures are not what the problems, once begun, are really about. In his own words now:

---

[21] This book consists of notes by Norman Malcolm of Wittgenstein's lectures in the spring of 1939. Unauthorized copies of these notes were made and put on sale in San Francisco in 1954. (The *Remarks on the Foundations of Mathematics* dates from 1937–1944.)

[22] Part I of the *Philosophical Investigations* was completed by 1945; Part II was written between 1947 and 1949.

[23] *Memoir of Wittgenstein* (London, Oxford University Press, 1958), pp. 53–54.

[24] Recorded in an unsigned mimeographed copy of the *Blue Book* in the Hoose Library of Philosophy at the University of Southern California.

> Certain verbal forms are misleading for a time and then cease to be misleading. The idea of Imaginery Numbers, e.g. There may have been something misleading in this when it was first introduced, but now it is utterly harmless; it just does not mislead anyone, while to say that the appellation "infinite development" is misleading is correct, though of course it does not mislead anyone in his calculations. It misleads in the idea of what they have done and the consequent interest which people have in certain calculations. The idea of the infinite as something huge does fascinate some people and their interest is due solely to that association, though they probably would not admit it. But that has nothing to do with their calculations. We could, e.g., play chess without the board or chessman, by description of the moves, etc. I might say chess would never have been invented apart from the board, figures, etc., and perhaps the connection with the movement of troops in battle. No one would have dreamed of inventing the game as played with pencil and paper. That would be right, but still the game (as played either way) would not be wrong. It is the same with mathematics. There is nothing right or wrong in a calculus, but it is its associations which make it seem worth while. But they are quite different from the calculus. Sometimes the associations are connected with practical applications; sometimes they are not. The association of the infinite is with something huge. Without this association no one would care a damn about the infinite.[25]

Now of course there is nothing wrong with words arousing pictures in our minds or with our tendency to associate the meaning of a word with the picture it arouses. It is different however with certain other words whose pictures tend to be highly misleading—the word "particle" for example. The pictures which such words tend to arouse in our minds lead us to expect the wrong things. Thus, for example, thinking of the formula "the cardinal number of all cardinal numbers" in terms of, say, "chairs" has a kind of dizzying effect on the mind, for the number involved is truly staggering. It conjures up, he says in *Math Notes*, "a picture of an *enormous colossal* number. And this picture has a charm." But the imagery here, although a natural and understandable consequence of our tendency to assimilate and correlate various expressions in our language, is entirely inappropriate. For, in fact, we have as yet, as he puts it, "no right to have an image. The imagery is connected with a different calculus: $30 \times 30 = 900$." The sense of amazement and puzzlement experienced at such occasions is simply a product of a misapplied image.

Wittgenstein goes further than that in fact. If something about a certain subject or problem "charms or astounds" us, he says, we may conclude from this that it is because we have been captivated by the wrong imagery. Imagery of that sort is a function of metaphors and such metaphors remain "fishy" as long as they are "exciting." When we begin to see these things in their true light, the amazement and excitement simply vanishes.

Thus, for example, certain parts of mathematics are regarded as "deep." "But the apparent depth comes from a wrong imagery. It is pedestrian as any calculus." Yet that is precisely the way people were misled about the infinitesimal calculus, when they mistakenly believed that it treated of infinitely small quantities. People said later, "If you look into it, there's nothing infinitely small there." But what did they expect to find; what would it be like for something to be infinitely small? Why are they disappointed? And what do they mean by saying it doesn't treat of the infinitely small? This might be contradicted. First, "infinitely small" hasn't a very familiar use. Second, instead of saying it doesn't treat of something infinitely small, we should say it doesn't treat of *small* things at all."[26] But it is because we think of such things in terms of such misleading images (in terms, for example, of sizes, as here) that we go so wrong. The amazement and excitement which such things inspire in us should be taken as a sign that we have simply been misled.

Wittgenstein speaks in these lectures of "charm," "excitement," "interest," etc., and it is easy to see how these expressions give way later on to such more familiar expressions from his published writings as, for example, "puzzlement," "wonder," "confusion," etc. The transition can be found even in these lecture notes themselves. For example, we find him remarking at one point:

---

[25] Lecture Thirteen, dated 5.6.35.

[26] *Math Notes*, pp. 62–63.

> There is one kind of misunderstanding which has a kind of charm. . . . we say that the line intersects at an imaginary point. This sets the mind in a whirl, and gives a pleasant feeling of paradox, e.g. saying that there are numbers bigger than infinity. . . . He has employed a sensational way of expressing what he has discovered, so that it looks like a different *kind* of discovery. . . . He describes a new state of affairs in old words, and so we don't understand him. The picture he makes does not lead us on. By the words of ordinary language we conjure up a familiar picture—but we need more than the right picture, we need to know the correct use.[27]

And that is precisely where such new notations, he goes on to emphasize here, fail us so badly. The fact is that "in an overwhelming number of cases people do have the same sort of images suggested by words. This is a mere matter of fact about what happens in our minds, but a fact of enormous importance."[28] In view of this, it is not difficult to see why and how confusions arise. For all that is really necessary for this to happen is for us to use familiar words in unfamiliar ways. The pictures aroused will be correct enough but, of course, they will be misleading. And it is in such misleading pictures, he concludes here, "that most of the problems of philosophy arise."[29]

In the *Blue Book* he puts this point in this way: "The new expression misleads us," he says, "by calling up pictures and analogies which make it difficult for us to go through with our conventions. And it is extremely difficult to discard these pictures unless we are constantly watchful."[30] We can be so watchful, he goes on to tell us there, by asking ourselves at such times, "*How far does the analogy between these uses go?*" We can also try to construct, he says, "new notations, in order to break the spell of those which we are accustomed to." These new notations, of course, later become his famous language-games. "The scrutiny of the grammar of a word," he says again further on, "weakens the position of certain fixed standards of our expression which had prevented us from seeing facts with unbiased eyes. Our investigation tried to remove this bias, which forces us to think that the facts *must* conform to certain pictures embedded in our language."[31]

We now turn, however, to the *Philosophical Investigations*. Difficulty lies, he explains it there, in our inability "to get away from the idea that using a sentence involves imagining something for every word." We fail to realize that we do all sorts of things with words—turning "them sometimes into one picture, sometimes into another."[32] Unfortunately, such pictures are often "only like an illustration to a story" and from it alone it is mostly impossible "to conclude anything at all"—for only "when one knows the story does one know the significance of the picture."[33] Mainly, however, the trouble with such pictures is that they seem "to fix the sense *unambiguously*" when this is not at all the case. On the contrary, "the actual use, compared with that suggested by the picture" is "muddied."[34]

Now certainly language has this effect on us—"the picture is there"; nor need we necessarily dispute its "validity in any particular case." But we do "want to understand the application of the picture."[35] And not only is this often lacking, but other pernicious effects result as well. Or, as he puts it later: "What this language primarily describes is a picture. What is to be done with the picture, how it is to be used, is still obscure. Quite clearly, however, it must be explored if we want to understand the sense of what we are saying. But the picture seems to spare us this work: it already points to a particular use. This is how it takes us in."[36]

In order to prevent ourselves from being taken in in this way we ought always to ask: does reality accord with such pictures?[37] These pictures *seem* "to determine what we have to do, what to look for, and how"—but they really do not do so. They *seem* "to make the sense of the expressions unmistakable" but in fact prove to be utterly misleading.[38] For example, "What am I believing in, when I believe that this substance contains two carbon rings? . . . there is a picture in the foreground, but the sense lies far in the background; this is, the application of the picture is not easy to survey."[39] In ordinary circumstances such words and the pictures which they generate "have an application with which we are familiar.—But if we suppose a case in which this application is absent we become as it were conscious for the first time of the nakedness of the words and the picture,"[40] of how "idle" such pictures are.[41] In the end we must simply regard them as "illustrated turns of speech,"[42] which stand "in the way of our seeing the use of the word as it is."[43] Philosophy, he concludes, "is a battle against the bewitchment of our intelligence by means of language";[44] "a simile that has been absorbed into the forms of our language produces a false appearance";[45] "a picture [holds] us captive."[46]

---

[27] *Ibid.*, p. 6.

[28] *Ibid.*, p. 39.

[29] *Ibid.*, p. 41. Cf. here *Philosophical Investigations*, I, 141.

[30] *Math Notes*, p. 23.

[31] *Ibid.*, p. 43.

[32] I, 449.

[33] I, 663.

[34] I, 426.

[35] I, 423.

[36] II, vii.

[37] I, 352.

[38] *Ibid.*

[39] I, 422.

V

I should not wish to suggest that one can open the *Foundations* and derive from it in some straightforward and immediate way the theory just expounded. This cannot be done even with the *Philosophical Investigations.* To derive a theory from any of Wittgenstein's writings one must pick and choose, order and rearrange, compress and expand. However, once a theory of this kind emerges the work does tend to fall into place. In connection with the *Foundations* some interesting statistics emerge. As it stands the book contains five parts and two appendices. These parts and appendices are divided into separate remarks of varying lengths, totalling 461 in all. Using as one's guide such key terms as "picture," "surprise," "interest," etc., one finds that more than a quarter of Wittgenstein's remarks are somehow connected with this theme. Although this should not be taken to mean that a fourth of Wittgenstein's book is devoted to this theme, it does mean that it is in the foreground of his thought and very much, obviously, permeated by it.

The textbook reproduces liberal portions from all the parts except the second Appendix. Although it does not therefore miss picking up some of the remarks connected with this theme, it picks up only about a fourth of them. This may not seem bad at all. But, unfortunately, it tends to miss the more helpful and illuminating remarks and, ironically, the single portion from which it fails to excerpt (the second Appendix) contains the most important material of all. Although this may have been intentional, the result of it is that the student who reads only the excerpts in this textbook finds himself at a disadvantage when confronted with the reviews, for the remarks which the textbook does pick up do not provide him with a sufficiently clear statement of this theme for him to be able to meet the challenge of the reviews. This is not to say that all the attacks upon Wittgenstein made by the reviewers can be met by appealing to this picture theory. Unfortunately that is not at all so. But as we shall see in a moment, some of the severe misunderstandings can be so clarified.

The theory which Wittgenstein expounds in these *Remarks* is essentially similar to the one which I have tried to reproduce above from other sources. Some of the points are even expressed in almost identical terms, although not always so explicitly as there. Remark number 17 in Appendix II, for example, is a case in point and should be compared with the similar observation (also dealing with the use of the word "infinite") reproduced earlier from *Math Notes.* It runs as follows:

> Ought the word "infinite" to be avoided in mathematics? Yes; where it appears to confer a meaning upon the calculus; instead of getting one from it. This way of talking: "But when one examines the calculus there is nothing infinite there" is of course clumsy—but it means; is it really necessary here to conjure up the picture of the infinite (of the enormously big)? And how is this picture connected with the *calculus*? For its connexion is not that of the picture / / / / with 4. To act as if one were disappointed to have found nothing infinite in the calculus is of course funny; but not to ask: what is the everyday employment of the word "infinite," which gives it its meaning for us; and what is its connexion with these mathematical calculi?

And in his next and final remark in this Appendix, he summarizes his results in these words:

> What I am doing is, not to shew that calculations are wrong, but to subject the *interest* of calculations to a test. I test e.g. the justification for still using the word . . . here. Or really, I keep on urging such an investigation. I shew that there is such an investigation and what there is to investigate there. Thus I must say, not: "We must not express ourselves like this", or "That is absurd", or "That is uninteresting", but: "Test the justification of this expression in this way". You cannot survey the justification of an expression

---

[40] I, 349.
[41] I, 291.
[42] I, 295.
[43] I, 305.
[44] I, 109.
[45] I, 112.
[46] I, 115.

> unless you survey its employment; which you cannot do by looking at some facet of its employment, say a picture attaching to it.

This remark is, of course, reminiscent of several other remarks quoted earlier from other sources dealing with this same theme. Throughout many of these remarks Wittgenstein is anxious to put us on our guard against our tendency to be bewitched by pictures which strike our fancy but which on closer inspection have no or extremely weak justification. As he puts it in remark number 16 here:

> If the interest here attaches to the proposition that has been proved, then it attaches to a picture which has an extremely weak justification, but which fascinates us by its queerness, like e.g. the picture of the "direction" of time. It makes one's thoughts reel mildly. Here I can only say: depart as quickly as possible from this picture, and see the interest of this calculation in its application.

That there may be nothing at all trivial in the point Wittgenstein is here urging may perhaps be gathered from a rather recent example—the recent attempts to make a landing on the moon. Scientists feared that such a landing would prove to be either difficult or impossible because of the thick layer of dust with which they believed the moon was covered. But the belief proved to be a myth. As one scientist connected with the project later remarked: it was an idea, he said, which people simply became fascinated with and would not give up, but there was really never any reason to entertain it at all. Wittgenstein has here obviously put his finger on some very deep and pervasive human tendencies which have far-reaching consequences.

In remarks 10, 11, and 12 he offers an extended example of the mechanics of such deceptions. Since his remarks here introduce us again to most of the key terms of the theory ("interest," "mystery," "picture," "excitement," etc.) I shall quote them in full before going on to summarize the main issues in my own words:

> 10. "Fractions cannot be arranged in an order of magnitude."—First and foremost, this sounds extremely interesting and remarkable. It sounds interesting in a quite different way from, say, a proposition of the differential calculus. The difference, I think, resides in the fact that *such* a proposition is easily associated with an application to physics, whereas *this* proposition belongs simply and solely to mathematics, seems to concern as it were the natural history of mathematical objects themselves. One would like to say of it e.g.: it introduces us to the mysteries of the mathematical world. *This* is the aspect against which I want to give a warning. When it looks as if . . ., we should look out.
>
> 11. When, on hearing the proposition that the fraction cannot be arranged in a series in order of magnitude, I form the picture of an unending row of things, and between each thing and its neighbour new things appear, and more new ones again between each of these things and its neighbour, and so on without end, then certainly there is something here to make one dizzy. But once we see that this picture, though very exciting, is all the same not appropriate; that we ought not to let ourselves be trapped by the words "series", "order", "exist", and others, we shall fall back on the *technique* of calculating fractions, about which there is no longer anything *queer*. The fact that in a technique of calculating fractions the expression "the next greatest fraction" has no sense, that we have not given it any sense, is nothing to marvel at. If we apply a technique of continuous interpolation of fractions, we shall not be willing to call any fraction the "next biggest".
>
> 12. To say that a technique is unlimited does *not* mean that it goes on without ever stopping—that it increases immeasurably; but that it lacks the institution of the end, that it is not finished off. As one can say of a sentence that it is not finished off if it has no period. Or of a playing-field that it is unlimited, when the rules of the game do not prescribe any boundaries—say by means of a line. For the point of a new technique is to supply us with a *new* picture, *a new form of expression*; and there is nothing so absurd as to try and describe this new schema, this new kind of scaffolding, by means of the old expressions.

I have been concentrating here entirely on the second Appendix but these relevant points and issues and their further elaboration are to be found scattered throughout the *Foundations*.[47]

Wittgenstein's doctrine of pictures, as some of these remarks show, has important and wide-reaching implications for the logic of discovery. Contained in them as well, however, is something that is of even wider significance. It has to do in a way, I believe, with the point long ago expressed by Bishop Butler that everything is what it is and not another thing. Part of the meaning of this remark is, I believe, that everything should be or ought to be evaluated in terms of the categories which it itself brings along. If we are looking at a work of art, for example, we should resist responding or evaluating it in terms of any external interests that it may have for us. A painting may be interesting to us because of the memories of our childhood it may awaken in us, or because an enormous amount of money was spent in purchasing it, or because it is a painting of the artist's son or mistress, but obviously none of these are what makes the painting a good painting—assuming it is a good painting. If these, then, are the reasons we find ourselves aroused, interested, and fascinated by the work, we can be certain that it is not to the painting as such that we are responding and that in a sense, therefore, we are misled by our *interest* in it.

---

[47] In addition to remark number i in this Appendix, the reader may find it valuable to examine the following remarks as well: *Part I:* Nos. 6, 14, 74, 118, and 125. *Part II:* No. 29. *Part III:* Nos. 12 and 21. *Part IV:* Nos. 6, 9, 16, 25, 29, and 37. *Part V:* Nos. 3 and 16.

The same may be said about our experience in connection with other matters as well. If, for example, we find ourselves responding with greater interest and enthusiasm to, say, literature than to mathematics, the chances are that our response to the former is of an external sort. For what may make the work of literature more interesting to us are perhaps the characters with which it deals and with whom we have a certain affinity, or their life stories which seem to shed some light on our own, or the generally large doses of sex which we find pleasurably arousing. But these in themselves are not what makes a great work of literature great and worthy of our attention. What makes it great is not its ability to give expression to these diverse, external needs but rather that it is able to exploit to the full its own peculiar medium and to give expression to itself. And in this sense both literature and mathematics—in fact all the arts and sciences—are as one. Each has its own goals and if they succeed in achieving them are as worthy and internally as interesting as any other. If therefore in view of this we find ourselves responding with greater interest to one as opposed to some other art or object, it is not because of anything that particular art or object (say, literature) possesses that the other does not. More likely, it is rather because we are responding to something that is not at all intimately connected with the thing as it is in itself. For from the point of view of the things themselves, they are simply what they are. If, then, one still tends to arouse more interest than some other object—or even, we might say, if one simply does arouse our *interest*—Wittgenstein is perhaps not wrong in urging us to become suspicious and enquire whether perhaps we are not looking at the thing through some misleading picture. Does infinitesimal calculus strike you as being "deep"? It is as "pedestrian" as any calculus, he would reply.

VI

Now that we have seen something of the wider background in terms of which Wittgenstein's remarks on mathematics must be read, we are perhaps in a better position to evaluate some of the criticisms contained in the three reviews before us. In looking at these reviews we shall not be interested here in evaluating their objections to Wittgenstein's handling of this or that particular mathematical doctrine—for that must be left to mathematicians to decide—but rather in seeing how far they have succeeded in grasping what we might call the intent and spirit of Wittgenstein's thought, particularly as that is revealed in his theory of pictures described above.

From this point of view, Michael Dummett's review is perhaps the least objectionable of the three. What seems to trouble him mainly is Wittgenstein's severe segregation of philosophy from mathematics. This is a pity, for it is precisely this failure on his part to realize fully Wittgenstein's reasons for doing so that has been responsible for Dummett's inability to appreciate the true scope of Wittgenstein's achievement here and from presenting a more balanced and fairer account of it.

Part of that achievement consists, as I tried to suggest a moment ago, in providing us with a way of talking about these matters which enables us to gain a firmer grasp of them. Dummett is not unaware of the power of this new way of talking. He points out, for example, that the opposing pictures used by Platonists and constructivists—the picture of our making discoveries within an already existing mathematical reality and the picture of our constructing mathematics as we go along—often lead their advocates to believe that everything else can be justified in these terms. They sometimes argue, he says, "as though it were a matter of first deciding which of these pictures is correct and then drawing conclusions from this decision" when "it is clear that these are only pictures," and "the dispute as to which is correct must find its substance elsewhere."[48]

This is, of course, good Wittgensteinian doctrine. Surprisingly, however, Dummett goes on to suggest that in his discussion of proof Wittgenstein himself has apparently fallen victim to this error. Although it is certainly not inconceivable that Wittgenstein should commit the very error which he is so anxious to warn others against, the fact that Dummett must resort to Wittgenstein's own terms to point this out, is of course a tribute to their value and power. For that is a kind of criticism which, after all, Wittgenstein himself was the first to make possible. If Dummett does not quite acknowledge this, the other reviewers are even less gracious about it, as we shall see in a moment.

---

[48] *The Philosophical Review*, *op. cit.*, p. 346.

Alan Ross Anderson's review although briefer than the other two is no less severe. Noting that "these fragments were written concurrently with Part I of the *Philosophical Investigations*, perhaps for inclusion in the latter work, had Wittgenstein lived to complete it," he finds it unsurprising "that there is little here that is fundamentally new." "Many of the problems discussed," he points out further, "have already been treated in the earlier volume, and though the remarks here are different in detail, they are the same in spirit." "What *is* new here," he notes however, "is the detailed discussion of several important results in the classical foundations of mathematics . . . and of the relation of logic to mathematics."[49] But these discussions too, he says, are merely further illustrations of Wittgenstein's general thesis, and whatever is substantive about them is unfortunately "directed to the battles of twenty-five to fifty years ago"—battle lines that are no longer drawn where they once were.[50] It is very doubtful, he therefore concludes, "whether this application of his method to questions in the foundations of mathematics will contribute substantially to his reputation as a philosopher."[51]

To support these doubts, Anderson proceeds to examine Wittgenstein's treatment of Gödel's theorem. But Anderson's treatment of this discussion is very curious. He realizes, for example, that a good deal of the point of Wittgenstein's discussion here is designed to change our *attitude* toward contradiction. But rather than concentrating on that, as he therefore ought to, he tries to show instead that Wittgenstein misunderstood the *facts* of the case under discussion. I shall not pause here to question whether it is possible to substantiate or sustain the kind of sweeping condemnation expressed in his concluding remarks by merely examining, as he does, one single, abstruse example. Anderson himself is uneasy about this and apologizes to the reader for it. Although his defence is that examples play an important role in Wittgenstein's method, he no doubt suspects that in view of the kind of method it is, no single example can really be considered decisive, one way or the other. Nor will I pause to question the facts of the case. It may be that Anderson's understanding and description of them are completely correct. But what is puzzling about his treatment is that when he does come to deal with these matters on the level on which they were designed to be dealt with, he comes, like Dummett, to charge Wittgenstein with offences which Wittgenstein himself was the first to make intelligible, and there is certainly something ironic about this—if indeed nothing worse.

Anderson grants, for example, that "it may be true that some writers, on the basis of a superficial familiarity with Gödel's result, tried to make something of a mystery of it," and that "it is no doubt in part against such writers that Wittgenstein is aiming his remarks." But, he goes on to declare, "the way to dispel the mystery is *not* to belittle the result, or to argue that it doesn't show what it seems to show." On the contrary, "the way to dispel the mystery *is* to see the result as a straightforward unprovability theorem"—a theorem that "is not even particularly surprising, once one gets over the initial shock."[52]

There are two questions which these comments raise in one's mind: one, Anderson uses here the language of mystery, surprise, and shock, obviously believing it to have a certain point. But the only point that language has is the one given it by Wittgenstein, for although the concept of surprise has been present in mathematics since the Pythagoreans, the particular emphasis given to it here (and the one apparently accepted by Anderson) is Wittgenstein's own. Anderson must also therefore believe that this language is appropriate and illuminating, although perhaps inappropriately applied in this particular case. But even that is perhaps doubtful, for (and here I come to my second point) if Anderson agrees, as again he seems to, that Wittgenstein's argument is not necessarily directed against mathematics or mathematicians but rather against those who have made a mystery of it (and therefore that it is the "results" as so understood that is at issue here) then he must also agree that the best way to dispel the mystery is indeed to try to belittle these "results" and to point out that they do not show what they seem to show. For how else are we to succeed in showing that the results are really "straightforward" (or "pedestrian" as Wittgenstein would put it)? There is a question here, of course, regarding whose "results" are at stake. But even if we were to grant in this case that it is Gödel's or Russell's results that Wittgenstein is talking about (and not the interpretation placed upon them by others), it is still obvious, as has been shown above, that certainly in a vast majority of cases it is not mathematical inventions as such but their understanding and interpretation (their "office") which is Wittgenstein's theme in this book. On Wittgenstein's behalf, therefore, we might very well wonder whether Anderson himself has not somehow missed the point here.

---

[49] *The Review of Metaphysics, op. cit.*, p. 446.

[50] *Ibid.*, p. 457.

[51] *Ibid.*, p. 458.

[52] *Ibid.*, p. 456.

From the point of view of this paper, Bernays' review is perhaps the most objectionable of the three. Bernays begins by quoting two remarks from Wittgenstein as evidence that, as he puts it, "jocularity of expression is not lacking with Wittgenstein."[53] These are: ". . . What I am doing is, not to show that calculations are wrong, but to subject the *interest* of calculations to a test" (p. 63, no. 18); and "My task is not to attack Russell's logic from *within*, but from without. That is to say, not to attack it mathematically—otherwise I should be doing mathematics—but to attack its position, its office. My task is not to talk about Gödel's proof, for example, but to by-pass it" (p. 174, no. 16).[54] It may be that Bernays does not mean to be critical when he characterizes these remarks as "jocular." In a sense they are of course that; but from what has been said above, it is clear, I think, that there is dead seriousness in them as well, for we now know what this "interest" is, and what this "within-without" business entails here.

Later on, Bernays says of Wittgenstein that although he tries to maintain a nominalist position, yet, oddly, we find him frequently appealing to facts of "our actual mental attitude." Bernays supports this charge of contradiction in Wittgenstein by the following sample passage:

> "In a certain sense it is not possible, therefore, to appeal to the meaning of signs in mathematics, because it is mathematics itself which first gives them their meaning. What is typical for the phenomenon about which I am speaking is that the *mysteriousness* about any mathematical concept is not *straight away* interpreted as a misconception, as a fallacy, but as something which is at all events not to be despised, which should perhaps even be respected. All that I can do is to show an easy escape from this obscurity and this glitter of concepts. It can be said, strangely enough, that there is so to speak a solid core in all these glistening concept-formations. And I should like to say that it is this which makes them into mathematical products."[55]

It is difficult to see where Wittgenstein has contradicted himself here, for although he is indeed referring here to two quite different conceptions of mathematics, he regards one of these conceptions as completely mistaken. That Bernays is not fully aware of the meaning of this passage and especially of such terms in it as "mysteriousness," "glitter," etc., seems to be indicated by the comment he makes upon it. "One may doubt," he says, "whether Wittgenstein has succeeded here in showing 'an easy escape from this obscurity'; one may be more inclined to suspect that here the obscurity and the 'mysteriousness' actually have their origin in the philosophical concept-formation, i.e., in the philosophical language used by Wittgenstein."[56] But how, one might ask, could Wittgenstein be accused here of falling victim to the 'mysterious' then the whole point of the passage is that we should be on our guard against falling prey to this very thing. It is doubtful whether Bernays and Wittgenstein are speaking about the same thing here.

A page later, Bernays says that "Wittgenstein's method aims at eliminating a very great deal. He sees himself in the part of the free-thinker combating superstition. The latter's goal, however, is freedom of the mind; whereas it is this very mind which Wittgenstein in many ways restricts, through a mental asceticism for the benefit of an irrationality whose goal is quite undetermined."[57] Although sweeping in its condemnation, this charge is completely unfounded, for although Wittgenstein's method is designed to eliminate a great deal, there is nothing indeterminate about what it desires to eliminate. On the contrary, as we have seen he is very clear and specific about what this is. Had Bernays been more clear himself as to the kind of fictions and superstitions which the method was designed to eliminate, he would hardly have regarded the attempt as either irrational or confining in its effects. For after all, to eliminate such things as are involved here is not to diminish but to expand the human mind.

---

[53] *Ratio*, *op. cit.*, p. 2.

[54] *Ibid.*

[55] *Ibid.*, p. 4.

[56] *Ibid.*

[57] *Ibid.*, p. 6.

Bernays ends his first section with an even more sweeping denunciation, involving now not only Wittgenstein but also all those who have fallen under his spell. He writes:

> In contrast to the assertive form of philosophical statement throughout the *Tractatus*, a mainly aporetical attitude prevails in the present book. There lies here, it is true, a danger for philosophical pedagogics, especially as Wittgenstein's philosophy exerts a strong attraction on the younger minds. The old Greek observation that philosophical contemplation frequently begins in philosophical astonishment today misleads many philosophers into holding the view that the cultivation of astonishment is in itself a philosophical achievement. One may certainly have one's doubts about the soundness of a method which requires young philosophers to be trained as it were in wondering.[58]

In terms of the role which such words as "shock," "surprise," and "astonishment" play, as we have seen, in Wittgenstein's work, this remark represents a gross misunderstanding of the whole motive of Wittgenstein's philosophy.[59] For far from being designed to nurture astonishment, the whole point of that philosophy is to understand its source and nature, and, if possible, to dispel it.

VII

To summarize: Wittgenstein's *Remarks on the Foundations of Mathematics* has been received with mixed feelings. Certainly the reason for this is the unfinished and incomplete nature of the material out of which it is composed. Since this publication is merely an incomplete selection from a larger body of unfinished jottings, we are neither in a position to determine in what state Wittgenstein himself might have published them had he had the desire and opportunity to do so, nor whether anything essential to our understanding of the selections published is to be found among the remaining jottings. Although nothing of interest can be said about what Wittgenstein might have done with these remarks had he been able to do so, a number of reasons lead one to believe that one or two things of importance which might have had an effect on our understanding of the work as a whole might indeed have been omitted. A discussion of one such theme—the interest, surprise, and shock which certain mental pictures generate in mathematics—shows this to be indeed the case.

University of Southern California

Received July 5, 1966

---

[58] *Ibid.*

[59] And is on a par with his earlier comment that Wittgenstein must be mistaken about language being a game and therefore about "language-games" because "our natural language has in no way the character of a game" (p. 19).

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 4, October 1967

II. FACTS: WHAT THEY ARE AND WHAT THEY ARE NOT

R. M. MARTIN

I

THE word "fact" is a seemingly harmless word in our philosophical vocabulary. We feel free to use it without subjecting it to analysis. The supposition is that there are no special difficulties with it, and therefore we should move on to more important matters. Facts are thought to be what true synthetic propositions or statements express. The facts are what we are said to believe when we believe truly. The facts are what we seek to have explained in empirical science. The totality of facts is everything that is the case. . . . And so on and on. These statements themselves, however, and their ilk, are far from clear and hence in need of explication. And insofar as particular philosophical positions depend upon them are these positions themselves suspect and needful of clarification. It will thus not be idle to go to the very roots of the matter and attempt to determine the ontic status of the things (if such there be) which "fact" is supposed to denote.

Russell set the stage (in 1918) when he contended that there are

> particulars and qualities and relations of various orders, a whole hierarchy of different sorts of simples, but all of them . . . have in their various ways some kind of reality that does not belong to anything else. The only other sort of object you come across in the world is what we call *facts*, and facts are the sort of thing that are asserted or denied by propositions, and are not properly entities at all in the same sense in which their constituents are. This is shown by the fact that you cannot name them. You can only deny, or assert, or consider them, but you cannot name them because they are not there to be named. . . .[1]

In general there are two fundamental ways in which one can speak of a given entity, either by naming it or at least by having it as a value for a variable. And ordinarily, if it is named it is then also a value for a variable. To be sure, there are derivative ways of referring to it, e.g., by describing it, but to describe a thing is to presuppose already that it is a value for a variable. Unless facts then are either nameable or values for variables, it is not clear how we can speak of them at all. Russell insists that we cannot name them. He is not clear as to whether they are admissible as values for variables. But a moment's reflection should convince us that on his view they are not. Russell's ontology, at least in the Logical Atomism papers, consists of "particulars and qualities and relations of various orders" in accord with type theory. Presumably a fact is not a particular, nor a quality nor a relation of any order, and hence cannot be a value for a variable. It is thus difficult to see how Russell can contend that facts are the "sort of thing you come across in the world." There is simply no room for them among the admissible furniture—or rather, there is no furniture for them in the admissible room.

We could tarry with Russell at length, for he returned to the subject of "facts" time and time again, but without, it is to be feared, arriving at a clear or well-developed conception.

II

We might hope to find a gain in clarity in Ramsey's "Facts and Propositions" of 1927.[2] This paper seems mistitled, however, for it is concerned primarily with belief, and with facts and propositions only in their connection with belief. Of course philosophical interest in facts and propositions goes hand in hand with the analysis of belief. Propositions have traditionally been thought to be the objects of belief and facts are what propositions, when true but not analytically so, express.

We shall not argue in the present paper against the traditional notion of proposition. But we shall not accept it either. Propositions are surely as obscure as facts. Hence it will not do to presuppose them in the analysis of the latter. The two go hand in hand, both in need of analysis. If we can succeed in straightening out fact-talk the suspicion is that we shall have gone a long way toward straightening out proposition-talk as well. Thus we shall give ear to no one who glibly identifies the two or regards one as a special case of the other, unless he goes on to analyze adequately one or the other in independent fashion.

---

[1] B. Russell, *Logic and Knowledge, Essays 1901–1950* (London, Allen and Unwin, 1956), p. 270.

[2] F. P. Ramsey, *The Foundations of Mathematics* (New York; Harcourt, Brace and Co., 1931), pp. 138 ff.

Ramsey agrees with Russell that "the fact that so and so" is not to be regarded as a name. He also argues that it is not to be regarded as a description. We might reasonably be inclined to regard a phrase such as "the death of Caesar" as a "description of an event." Such phrases should surely be analyzed and given an exact logic. The first thing to note about them, however, is that they are not Russellian descriptions of the form "the one so-and-so." Event-descriptions are something quite different. Ramsey tells us singularly little about them. Strictly we should worry here as to what events are, as well as event-descriptions. Whatever events and event-descriptions may turn out to be, Ramsey insists that they are not to be confused with facts and fact-descriptions.

The connection between the *event* which was the death of Caesar, Ramsey writes,

> and the *fact* that Caesar died [italics added] is, in my opinion, this: "That Caesar died" is really an existential proposition, asserting the existence of an event of a certain sort, thus resembling "Italy has a King," which asserts the existence of a man of a certain sort. The event which is of that sort is called "the death of Caesar" [quotes added], and should no more be confused with the fact that Caesar died than the King of Italy should be confused with the fact that Italy has a King.[3]

Perhaps so, but we cannot be very clear about this contention until we are much clearer as to whether existential propositions "really" assert "the existence of an event of a certain sort." Are these propositions of the form "There exists an event E such that —E—," where "—E—" is a sentential function containing 'E' as its only free variable, 'E' ranging over events? Or are they of the form, "Such and such a kind of an event exists," using "exists" here essentially as a predicate? And in either case, an ontology of events seems presupposed, and this very ontology itself is in need of analysis. Nonetheless, this passage is of interest in suggesting an important interconnection between fact- and event-talk.

Ramsey notes[4] that, because phrases beginning "the fact that" are neither names nor descriptions, propositions "about 'the fact that aRb' must be analyzed into (1) the proposition aRb, (2) some further proposition about a, R, b, and other things; and an analysis of cognition in terms of relations to fact cannot be accepted as ultimate." In accord with this, "X believes that aRb" is to be analyzed in terms of some polyadic relation having as arguments the person X, the proposition aRb, the thing a, the relation R, the thing b, and "other things" (we are not told what), perhaps a time t.[5] The suggestion that belief-sentences be so analyzed is of the highest importance, but it is to be doubted that it throws much light on "fact" or the ontic status of whatever it is supposed to denote (if at all).

III

A notion of fact—we say 'a' not "the," for we do not know that there is just one—seems central to the ontology of Wittgenstein's *Tractatus*.[6] "The world is the totality of facts, not of things . . . The facts in logical space are the world . . . An atomic fact is a combination of objects (entities, things). It is essential to a thing that it be a constituent part of an atomic fact. . . ." These oft-quoted dicta do not, however, lead us to a clear notion of fact where others have failed. At best, as Max Black has suggested, "Wittgenstein has deliberately parted from the ordinary uses of "fact" without providing instructions for the new uses."[7] If exact instructions had been given, these dicta would perhaps have lost much of their depth and charm. A main advantage of clear philosophical writing, even if lacking charm, is that errors then become pellucid, whereas in the murky waters there is often only a seeming depth.

We need not comment at length upon the niceties of Wittgenstein's use of "fact" in the *Tractatus*.

---

[3] *Ibid.*, p. 141.

[4] *Ibid.*, pp. 141–142.

[5] Cf. the author's "Toward an Extensional Logic of Belief," *The Journal of Philosophy*, vol. 59 (1962), pp. 169–172; "On Knowing, Believing, Thinking," *ibid.*, pp. 586–600; and *Belief, Intension, and Ontology*, in preparation.

[6] L. Wittgenstein, *Tractatus Logico-Philosophicus* (New York; Harcourt, Brace and Co., 1922).

[7] Max Black, *A Companion to Wittgenstein's Tractatus* (Ithaca, Cornell University Press, 1964), p. 34.

Others have done this ably. Instead, we point out certain dicta for future reference. Especially interesting for such purposes are 2.03–2.032: "In the atomic facts objects hang one in another, like the members of a chain. In the atomic fact the objects are combined in a definite way. The way in which objects hang together in the atomic fact is the structure of the atomic fact."

IV

Carnap, in *Meaning and Necessity*[8] is inclined to identify facts with a certain kind of proposition, taking propositions now in the "objective sense" as intensions of sentences. A proposition is thus (p. 27) not "a linguistic expression nor . . . a subjective, mental occurrence, but . . . something objective that may or may not be exemplified in nature." A fact for Carnap is not just any proposition, but rather one which is true, is factual or synthetic, and is "specific or complete in a certain sense." That a proposition (in Carnap's sense) should be both true and factual in order to be a fact is evident. False propositions clearly are not, nor are there facts corresponding to analytic or L-true propositions. The requirement of specificity is not so clear, however, and Carnap is less certain of it anyhow.

Carnap considers a true factual proposition p that a certain piece of paper is blue. "But the property Blue," he says, "has a wide range; it is not specific but includes many different shades of blue, say Blue$_1$, Blue$_2$, etc." The given piece of paper, however, has only one of these shades, say Blue$_5$. The proposition q that it is Blue$_5$ is more specific than the proposition p that it is Blue. Hence, Carnap thinks, p should perhaps not be regarded as a fact. But even Blue$_5$ may be further subdivided. If we suppose that it is not, however, q still may lack specificity with respect to other properties, e.g., size, shape, etc. If we go on in this way, Carnap suggests, and "require of a fact . . . [the] maximum degree of completeness (short of L-falsity), then there is only one fact, the totality of the actual world, past, present, and future."

There are several difficulties with this proposal, some of which no doubt Carnap would admit. For one thing, if there is only one fact, the sum total of the universal history of the actual world, there is not then much use for the term "fact" at all. Further, we do use the word in the plural and such uses should surely be accommodated. There is also another matter, more important, which it is curious that Carnap does not discuss. When we speak of the "maximum degree of specificity," we should really add: relative to the language-system being used. To speak of facts is to speak *in* a certain language-system. There seems to be just no way of getting around this circumstance at all. Hence we can speak of a maximum degree of specificity only relative to the language-system at hand, and hence only relative to its vocabulary. We can never hope to specify *all* the properties which the given piece of paper has. The best we can do is to specify those for which the language-system at hand has expressions. Of course Carnap is speaking of propositions, not of sentences, and propositions are outside language. We get at propositions in his sense, however, by means of language, more particularly by means of the sentences which have them as their intensions, as Carnap puts it. The best we can hope for, in trying to indicate the degree of specificity of a fact in Carnap's sense, is by means of some declarative sentence in some language-system. The notion of a maximum degree of specificity independent of language seems a myth.

Carnap agrees that there are enormous difficulties to be overcome before we can achieve a satisfactory notion of proposition. He is hopeful that this may be done, but only within an intensionalist kind of semantics. Formidable arguments, however, can be mustered up against all forms of intensionalism, or at least against all forms which admit primitively intensions *sui generis* as values for variables.[9] Thus it is doubtful that a satisfactory theory of propositions, and hence of facts, along Carnap's lines can be achieved.

V

In their Bristol symposium on "Truth" in 1950, Austin and Strawson agree, for differing reasons, that it would be wrong to identify "fact" and "true statement."[10] Concerning "fact" Austin is a bit cavalier perhaps, and no clear ontology for it is suggested. Strawson thinks that Austin fails to discriminate sufficiently "thing," "event," "situation," "state of affairs," "feature," and "fact." "All these are words," Strawson urges, "which should be handled with care." Surely—but let us go on to search for an exact (or even inexact) denotation for "facts" in Strawson's comments. He writes:

---

[8] (Chicago, University of Chicago Press, 1947), p. 28.

[9] See the author's *Intension and Decision* (Englewood Cliffs, Prentice-Hall, 1963), and "An Improvement in the Theory of Intensions," *Philosophical Studies* (to appear).

[10] "Pyschical Research, Ethics and Logic," *Proceedings of the Aristotelian Society*, Supplementary Volume XXIV (London, Harrison and Sons, 1950), pp. 111–172.

> Facts are known, stated, learnt, forgotten, overlooked, commented on, communicated or noticed. . . Facts are what statements (when true) state; they are not what statements are about. They are not, like things or happenings on the face of the globe, witnessed or heard or seen, broken or overturned, interrupted or prolonged, kicked, destroyed, mended, or noisy.[11]

This interesting passage suggests that facts may have certain epistemic properties, but otherwise it proceeds on the *via negativa*. But we are no better off than we were before. It is to be feared that we cannot be very clear about "Person X states fact f," "Person X learns fact f," etc., etc., without some indication of what the f's are. And similarly, if facts are what true statements "state," we are left presumably with such forms as

True statement a states the fact f,

or perhaps even

True statement a states the facts $f_1, \ldots, f_m$,

dangling rather precariously without analysis. Resolute Oxford laymanship, to paraphrase Quine, presumably prevents bringing in strict logico-semantical or structural-linguistic considerations.

Perhaps Strawson regards facts as a kind of pseudo-entity. But even pseudo-entities, e.g., virtual classes, have their proper dignity.[12] If they are merely pseudo-entities, to be discussed in a *manière de parler*, it is truly "unfair to facts" not to give them their proper ontic status.

Are facts "things in the world," or not? Austin thinks so, but Strawson not. Strawson is worried primarily about "fact" as "wedded to that-clauses," whereas Austin contends that "to explain the meaning of 'fact' in terms of the expression 'fact that' is to invert the real order of things. . . ."[13] Curiously, Austin's "real" order turns out to be essentially the etymological order as given in the *O.E.D.* Originally "fact" stood for "something in the world" but later becomes embodied in "fact that" as a "grammatical convenience." But the "real order," whatever it is, is surely not to be identified with the etymological order. It might happen that they coincide, but one would not know this in advance. It is to be feared then that Austin has not made his point, committing here perhaps some form of genetic fallacy.

It is surely a good admonition that we should "handle with care" such words as "thing," "event," "situation," "state of affairs," "feature," and "fact." Possibly some of these might turn out to be equivalent to each other, in certain uses at least, but possibly not. Unfortunately neither Austin nor Strawson has given us proper instructions for the careful handling of these fragile words.

We cannot tarry further with the quarrel between Austin and Strawson. That was their little affair anyhow and is now a thing of the past. Since neither looked beyond language to logic or meta-logic,[14] it would be miraculous if either of their views could have turned out satisfactorily. Without a proper logical underpinning, it would seem, there is not much hope for success in philosophic analysis. The use, implicit if not explicit, of a proper logic and meta-logic gives a necessary condition surely for adequacy of analysis, although of course by no means a sufficient one.

VI

A more sophisticated discussion of "fact" is to be found in Scheffler's *The Anatomy of Inquiry*.[15] Attention is confined to contexts of the form "explains-the-fact-that." Scheffler, however, has not wedded facts to "fact-that" clauses, for he takes full note of what is involved in the admission of a fact ontology. Such an admission he finds objectionable and suggests two ways of handling "explains-the-fact-that" which involve no such ontology.

If we say, as seems natural, that the aim of a scientific explanation is "to explain the facts," and seek to determine precisely, at the meta-logical level, what such explanations are, we seem driven to admitting facts in some guise as explananda. The object of explanation is then not a "thing." Scheffler wrote that it is:

---

[11] *Ibid.*, p. 133.

[12] See the author's "The Philosophic Import of Virtual Classes," *The Journal of Philosophy*, vol. 61 (1964), pp. 377–387.

[13] J. Austin, *Philosophical Papers* (Oxford, Clarendon Press, 1961), p. 112.

[14] Cf. W. V. Quine, *Word and Object* (New York and London, John Wiley and Sons, 1960), pp. 270 ff.

[15] (New York, Alfred A. Knopf, 1963), pp. 57 ff.

> no longer the spatio-temporal individual chunk, but something of another sort associated with it, of which there are as many as there are logically independent descriptions of the chunk. These new entities (let us call them hereafter "facts"), are not themselves spatio-temporal entities: they are neither dated nor bounded. Nor are they identified with the descriptions themselves. They are abstract ("logically intensional" [in Quine's phrase]) entities, intermediate between chunk and descriptions, each such entity corresponding to some class of logically equivalent (true) descriptions uniquely.[16]

Scheffler does not tarry long with this "picture" of facts but moves straightaway to avoid the dubious ontology involved. He suggests two methods. One requires fundamental use of "EF" (for "explains-the-fact-that"), the problem being to assure its proper syntactic behavior. Let "*Bk*" express that the spatio-temporal chunk *k* is blue, and let *g* be a suitable class of premiss-statements. (It is perhaps extraordinary for Scheffler, in the wake of Goodman's nominalism, to speak of a "class" of premisses. He should rather speak of a *virtual* class here or of a finitely long conjunction.) How, then, does the meta-linguistic sign "EF" behave? Scheffler does not bring out clearly enough that he is operating within quite a complicated meta-language, and hence the reiteration here of "meta-linguistic." He writes that the meta-linguistic expression

$$(1)\ g\ \mathrm{EF}(Bk)$$

> is to be taken as consisting of the name '*g*' and the one-place predicate "EF(*Bk*)," formed by inserting "*Bk*" into the second blank of ". . . EF – – –." The whole statement thus applies a one-place predicate to the sentence-string [sic!] *g*. It does not relate *g* to anything else. . . .

A more perspicuous notation would be

$$(2)\ \mathrm{EF}(Bk)\ g$$

wherein, as in "*Bk*" itself, the predicate is given first. If '*g*' is the name of a finite conjunction of premisses, (1) or (2) seem acceptable locutions. But if '*g*' is a class word, which (on p. 66) Scheffler explicitly says it is, then (1) or (2) involves an abstract ontology, albeit a different one from that which he is trying to avoid. So regarded, (1) or (2) surely cannot be regarded as acceptable locutions for a presumed nominalist such as Scheffler.

We must now query the status of "EF—," or rather of the "EF—"s, for there is a vast multiplicity of them, within the whole meta-logical system incorporating the desired explanation-talk. Are they *primitive* predicates, each and every one? If so, what an extraordinarily excessive array, a different one needed for each different explanation! If they are not primitives, how are they defined? Some somehow in terms of others? Scheffler gives no hint of an answer. We are told, however, that "*g* EF(*Bk*)" is true if and only if "*g* fulfills the criteria of explanation with reference to *Bk*." Now we may well ask: What kind of a suggestion is this supposed to be? Is it a logical consequence of a definition of "true in the meta-language"? At any event, the statement itself appears to be one in the meta-meta-language and hence cannot be an axiom of the meta-language incorporating the explanation-talk. The matter is of some little consequence, for unless we are given axioms governing the EF(—)'s, or unless they are characterized in some suitable alternative way, we are not told what the supposed theory concerning them really is.

Scheffler goes on to suggest a second method in order "to avoid the merest whisper of reference to facts." Here a sign 'W' is introduced as a "predicate-forming operator, producing a *one-place predicate of sentences* when its blank is filled by a sentence." Thus "W(*Bk*)" is a one-place predicate applicable to sentences, and to say that a sentence is a W(*Bk*) is to say that that sentence is a why-*k*-is-blue sentence. The why-*k*-is-blue sentences are just those logically equivalent to "*Bk*." Scheffler had earlier suggested that the fact-that-*k*-is-blue might be regarded as the class of sentences logically equivalent to "*Bk*," but had rejected it because of the abstract ontology involved. (The *virtual* class of sentences logically equivalent to "*Bk*," Scheffler fails to note, is not objectional in this way.) We gain the effect of talking of facts in essentially this sense, however, "if we construe 'W(*Bk*)' is a predicate applying to every logical equivalent of the sentence within the parentheses, inclusive of this sentence itself." Finally, to render this whole discussion "explicitly concrete," Scheffler regards sentences, and indeed expressions generally, as *inscriptions* or expression-events rather than abstract shapes or expression-designs as is more customary.

Scheffler suggests no axioms concerning the "W(—)"s, just as he had suggested none earlier concerning the "EF(—)"s. Those needed for the "W(—)"s would have to provide a kind of inscriptional account of logical or L-equivalence. It is a pity that Scheffler does not say anything about this, for no such account seems ever to have been given.[17] If we presuppose a syntax based on expression-shapes, a syntax of the classical Carnap-Tarski kind, we may go on to a notion of L-truth or analyticity, using Platonistic devices, and thence to L-equivalence. To avoid this, and in order to presuppose only a syntax of inscriptions, Scheffler in effect takes L-equivalence (a special kind of analyticity) as a primitive in the form of a vast multiplicity of "W(—)"s. Whether this is a wise, or even really honest, procedure is to be doubted. It is just packing too much up into those "W(—)"s which needs unpacking and analysis.

---

[16] *Ibid.*, p. 59.

In the second method (that involving the "W(—)"s) we have lost all trace of reference to facts, it is true. But in the first method too there is only the faintest lingering aroma of them, merely in the choice of 'F' in "EF." Scheffler could equally well have chosen 'T' for "that" and written "ET" for "explains that." We would not think we had therewith gained much illumination concerning a realm of *thats*. Of course the thats are eliminable, just as in his account the facts are. And this no doubt is as it should be.

In the second method, that employing the "W(—)"s, we seem to have lost all connection with explanation-talk. Although the 'W' is supposed to be read "why," it is not clear that it should be, due to the use made of L-equivalence. That a sentence is L-equivalent to another is not in general a ground for regarding it as giving an answer to the question "Why?" In this second method we have lost sight of the explanatory premiss-string g. Further, it has often been suggested that a proposition, as we have noted above, may be regarded as a class or virtual class of L-equivalent sentences. Now "W(—)" in effect functions as standing for just such a class. Hence Scheffler is in effect identifying facts with propositions so regarded, surely an illicit identification.

VII

Although they are presumably eliminable, we need not conclude therewith that facts are uninteresting. Russellian descriptive phrases are eliminable, but both interesting and worth defining. The question arises: Can we construct a "model" for facts, handling them not as an Austinian "something in the world," but rather somehow as fictitious intensional entities, spoken of only in a *manière de parler* and within a proper kind of meta-language in which the main uses of "fact" may be accommodated? Let us attempt to answer this in the positive. But first a word about intensional entities, a word about the *manière de parler*, and a word about meta-linguistic propriety.

That facts are intensional entities of some kind has perhaps not always been fully recognized. Of course, if they are identified with some sort of proposition, their intensional status comes out clearly. Even if propositions are not available, as here, we must still recognize the intensional status of facts. Scheffler brings this out conclusively.[18]

The device of using a *manière de parler* is a familiar one since Russell's first writings on descriptions. Russell succeeded, it will be recalled, in defining in context phrases of the form "the one and only x such that —x—," where in place of "—x—" we put in some sentential function containing 'x' as its only free variable. Such phrases are regarded as *manières de parler* within the wider sentential contexts in which they occur. They are not defined in isolation, but only within the wider contexts. Another use of the device of the *manière de parler* is in the notation for virtual classes.[19] Virtual classes are not included among the values for variables, however, whereas the one so-and-so, as introduced by Russell, is actually one of the objects so taken. There is this important difference between the two uses. But there is similarity in that the expressions for virtual classes also are defined only within the broader sentential contexts in which they occur.

Elsewhere semantical meta-languages based on denotation, for first-order-object-languages, have been characterized in some detail, and their presumed adequacy for philosophic analysis and the philosophy of science established. Also a strong theory of the many kinds of intension may be built up within such a meta-language.[20] Intensions turn out to be virtual entities of various kinds. Strictly there are no such things as either virtual classes or intensions. But this does not require that we cannot speak of them, or that such talk is unimportant or useless. Utility or interest is not to be confused with being the value for a variable.

---

[17] It is difficult enough to give a satisfactory account of L-equivalence within a first-order classical semantics, let alone an inscriptional one. Cf. the author's *The Notion of Analytic Truth* (Philadelphia, University of Pennsylvania Press, 1959).

[18] We cannot improve upon his presentation, and therefore refer the reader to pages 66–68 of his book.

[19] See "The Philosophic Import of Virtual Classes," *op. cit.*

[20] See *Intension and Decision* and "An Improvement," *op. cit.*

Let us now go on to attempt to construct a "model" for facts within the meta-languages based on denotation. Of course certain heuristic principles must guide any such construction. Certain features which we wish the "facts" to have must be selected, and others rejected. And concerning any such selection there is bound to be dispute. It is not clear that any one selection will give us the one and only one true picture of the "facts." Rather we must put forward a view and try to make it as adequate as we can. If we fail, there is then at least the possibility that it can be improved or rendered more adequate in this respect or that. But one thing surely is clear at this late date: we cannot with impunity philosophize about facts in resolute laymanship and in disregard of the refined tools of modern semantics. To do so is to rely upon mere horse and buggy procedures, and to get only to the next town when we wish to go a goodly distance.

A few features concerning "facts" which the theory here will incorporate are the following.

(i) We shall speak of facts only relative to a language-system *L*. It is fact-*talk* that interests us, and facts only insofar as we can have fact-talk about them. Given any *L*, there are no doubt always facts beyond the reach of the modes of expression of *L*. Of such facts there can then be no fact-talk within *L*. Presumably we can always move on to a richer system *L′* which takes cognizance of these additional facts. Presumably also there is no richest system in which all facts, in some sense, can be accommodated. The very conception of such a system presupposes a notion of *all facts* independent of language. It is very doubtful whether such a notion can be given any strict meaning. However this may be, we need here the relativization of fact-talk to a specific and well-articulated *L* regarded as object-language.

(ii) Facts are to be discussed only within a semantical meta-language. In their analysis, there is reference not only to the things of the object-language but also to the expressions of that language. Now semantics is the theory of how such things are brought into relation with such expressions. Failure to realize the full semantic structure of facts has no doubt impeded previous writers on the subject. Perhaps an adequate notion of fact can be built up in some other than a semantical way, but this is doubtful, and at any event seems never to have been done.

(iii) Facts are eliminable in the sense that expressions for them are. More specifically, any meta-linguistic sentence containing "fact" is by definition a shorthand for a longer sentence, in the primitive notation of the meta-language, in which neither that word nor any equivalent occurs. That facts are eliminable does not make them uninteresting or not worth bothering about, as we have already in effect remarked. The cardinal number two is eliminable (in the same sense) in the Zermelo set theory, for example, but we do not conclude from this that that number is somehow uninteresting or unimportant, or that there is then something the matter with that set theory. (There may well be something the matter with that theory, but not for this reason.)

(iv) Facts are "fictitious" entities in the sense that they are in no way values for variables within the meta-language—or indeed in any language. Nor are they "named" in any sense by the primitive notation. Of course, expressions for them can be built up in context. But such expressions are not names in any proper or strict sense. Just what kind or kinds of expressions are needed we shall note as we go on.

(v) Facts are most intimately bound up with the notation of *L*. In a certain sense, every meta-linguistic sentence containing reference to a fact is reducible to a sentence describing the notation for that fact. Facts are like what elsewhere have been called *quasi-syntactic* or *nominal* virtual classes.[21] Although we speak of them *as if* they were legitimate objects, strictly our sentences refer instead to the notation.

(vi) Facts are intimately linked with the semantic notion of truth. Surely in the definiens of the definition of a sentence containing "fact," "true" will occur in a fundamental way. No matter what facts are, they are surely not independent of truth, and the notion of truth seems most efficiently handled in terms of the modern semantic notion. Many analytic philosophers think they can avoid bringing in the semantic concept, and that some rather vague, pre-analytic notion will serve equally well. Perhaps this was so back in the horse-and-buggy days. The view here is that the semantic notion cannot be avoided, and the sooner we become accustomed to this the better for us all.

(vii) Facts are in some sense "logically intensional." This does not entail, however, that the meta-language in which our fact-talk is incorporated is in any way an intensional meta-language in the objectionable sense of having intensions *sui generis* as values for variables. The meta-language rather is of the usual extensional kind, as we have already noted. Within it, a notion of logical or *analytic truth* may be defined, and hence also the notion of *L-equivalence*. Whether or not we decide to follow Scheffler in making fundamental use of L-equivalence, we shall surely need the notion of logical truth. But all such notions, notions of what Carnap calls *L-semantics*, can be adequately accommodated within the purely extensional meta-languages. Likewise we shall use a notion of *L-designation*, about which more will be said as we go on.

---

[21] In "An Improvement," *op. cit.*

(viii) Facts have an internal structure. In them objects are "combined in a definite way," this way being "the structure of the . . . fact." We can, if we like, distinguish between "atomic" and "molecular" facts, but not independently of the given language-system L. This distinction thus loses the intrinsic ontic flavor that it had for Russell and Wittgenstein, and is hence here unimportant. The structure of facts, however, is significant, but again, is not described ontically but rather in terms of the expressions used. The structure, in short, is a linguistic structure rather than an ontic one.

(ix) Facts are concerned primarily with particular individuals, ascribing to them, as it were, a given property or relation. *General* facts, if there are such, are then to be mere congeries in some sense of facts about particulars. One might wish to condone general facts in some other sense, but this does not seem needed and at any event will not be done here.

(x) Facts are intimately linked with the semantic notion of *factual* or *synthetic* truth, not just with that of truth. Nonetheless facts are not just factual truths. They are a kind of pseudo-entity, so we shall maintain, sharply to be distinguished from linguistic expressions. But more of this anon.

VIII

In his *Word and Object*, Quine has extolled the notion of an ordered pair (or unit relation) as a philosophical paradigm.[22] In any situation in which we have two entities in a given order and wish to consider them as one, the device of using the ordered pair recommends itself. Quine concerns himself with *real* ordered pairs in the sense that they are themselves values for variables. He writes: "A notion of ordered pair would fail of all purpose without ordered pairs as values of the variables of quantification."[23] But this is too strong. We can put *virtual* ordered pairs to work and get good service out of them. Likewise virtual ordered triples, etc., should we need them. Facts are here to be regarded as virtual ordered pairs, triples, etc., of a highly special kind.

Let us presuppose now the theory of virtual classes for some first-order language-system L. Further, we presuppose a semantics for L, of the usual kind, based on denotation, with notions of truth and analyticity explicitly defined.

Let us introduce now, first *virtual-class designation* and then virtual-class *L-designation*. We say that an expression a *designates* the virtual class $x\ni(\text{—}x\text{—})$ if and only if a is a one-place predicate constant and for all x, a denotes x if and only if $\text{—}x\text{—}$, where "$\text{—}x\text{—}$" is a sentential function of L (or translation of such) containing 'x' as its only free variable. We can go on to say then that a *L-designates* $x\ni(\text{—}x\text{—})$ provided a designates $x\ni(\text{—}x\text{—})$ and $a=\mathbf{m}$, where in place of '$\mathbf{m}$' we put in the (structural-descriptive) name of the abstract "$x\ni(\text{—}x\text{—})$." However, we may drop the clause concerning designation here, because it follows from the Rules of Denotation presupposed.[24] Hence a L-designates $x\ni(\text{—}x\text{—})$ if and only if $a=\mathbf{m}$, where (etc.)

In this last definition the specific notation used to refer to the virtual class $x\ni(\text{—}x\text{—})$ is important, for '$\mathbf{m}$' must be taken as just the (structural descriptive) name of "$x\ni(\text{—}x\text{—})$." L-designation is, strictly speaking and as construed here, a notion of syntax, not of semantics. Hence the virtual classes referred to in the definiendum of the definition of "L-designates" are called *quasi-syntactic* or *nominal* virtual classes.

We note then that the relation of L-designation here differs remarkably from that of Carnap. That of Carnap is of course a semantical relation, not one of syntax—and further, it seems to be one within the theory of meaning or intension rather than within the theory of reference. According to him, and in essentially his own words, an expression a L-designates an entity in L if and only if it can be shown that a designates that entity merely by using the semantical rules of L without any reference to facts.[25] For Carnap, the entity L-designated may be a real class, whereas here it can be a virtual one only. On the other hand, suppose we know that $a = \mathbf{m}$, where '**m**' is the structural description of "$x \ni (\text{—}x\text{—})$." It then follows, merely by the Rules of Denotation, that a designates $x \ni (\text{—}x\text{—})$, as we have already remarked. This circumstance seems sufficient to justify calling the relation here one of L-designation.

---

[22] Pp. 257 ff.

[23] *Ibid.*, p. 258.

[24] Cf. the author's *Truth and Denotation* (Chicago, University of Chicago Press, 1958), pp. 108 ff.

[25] *Meaning and Necessity*, *op. cit.*, p. 163.

We can also speak of the *L-designation of individuals* as follows. Here likewise we speak of nominal, rather than of real, individuals. Let the roman letters 'a', 'b', etc., be primitive individual constants. We then wish to affirm that 'a' L-designates the nominal individual a, 'b' the nominal individual b, and so on. To make this exact, let us use (as we have been doing) the italic letters 'a', 'b', etc., as variables of the meta-language ranging over the expressions of L. We introduce, in the meta-language, an expression "*ay*" as the (structural-descriptive) name of 'a', "*bee*" of 'b', and so on. Then in general, we may let a LDes **a** abbreviate $a = \mathbf{ay}$ where in place of 'a' we put in an individual, constant and in place of "**ay**" its structural descriptive name. Hence a LDes a if and only if $a =$ 'a', i.e., $a = ay$, a LDes b if and only if $a =$ 'b', i.e., $a = bee$, and so on.

Let us write "$<x,y>$" for the virtual ordered couple the first member of which is x, the second y. And let us write

(1) $<x, y \ni (\text{—}y\text{—})>$

for the virtual ordered couple, the first member of which is x and the second the virtual class $y \ni (\text{—}y\text{—})$. Expressions of the form (1) provide a general notation for facts, when embedded in the proper contexts.

Let us use

(2) $<\mathbf{a}, x \ni (\text{—}x\text{—})>$-is-a-fact

as short for the following:

> There are expressions a and b such that a LDes **a**, b LDes $x \ni (\text{—}x\text{—})$, and the result of writing b followed by a is true but not analytically so and is L-equivalent to no sentence not containing **a**,

where in place of '**a**' we put in an individual constant.

We note first that we have here a definition-*schema*, not a definition. The schema stipulates an infinity (strictly, a double infinity) of definitions, one for each specific choice for '**a**' or "$x \ni (\text{—}x\text{—})$." Each instance of the schema we may think of as providing a definition.

Suppose the nominal individual a is a member of the nominal virtual class $x \ni (\text{—}x\text{—})$, truly but not analytically so. Then the virtual ordered pair

(3) $<\mathrm{a}, x \ni (\text{—}x\text{—})>$

is a fact according to this definition. To say this, however, is rather misleading. For it suggests that the subjects of the ordered pair are, respectively, the nominal individual a and the nominal virtual class $x \ni (\text{—}x\text{—})$. Also it suggests that we have the ordered pair floating around somehow independently of its being talked about in the particular context in which it is declared to be a fact. Both of these suggestions are erroneous.

Strictly the whole expression

(4) $<\mathbf{a}, x \ni (\text{—}x\text{—})>$-is-a-fact

is defined as a unity and cannot be broken up. Strictly there is no ordered couple here being talked about. Nor are we talking about the individual a or the virtual class $x \ni (\text{—}x\text{—})$. Rather, as we see by glancing at the appropriate definiens, we are speaking of the *expression* 'a' and of the *expression* "$x \ni (\text{—}x\text{—})$," ordered in a certain way. In the definiendum we are speaking of the individual a and of the virtual class $x \ni (\text{—}x\text{—})$ only nominally, as we have noted. Although we gain the effect there of speaking of them really, this is merely an illusion.

When we speak nominally of an individual a we are only one step removed from a, as it were. We seem to be speaking of the individual a but actually are speaking of an expression which L-designates a. When we are speaking quasi-syntactically of a virtual class, however, we are two steps removed. The virtual class $x \ni (\text{—}x\text{—})$ itself is one, the expression which L-designates it, another. A third stage of removal is provided by the ordering. We cannot correctly say that facts are virtual ordered pairs, for the order is wholly embedded in the whole context of the form (4). We have referred to them above as such ordered pairs, but this is not strictly proper. We might call them *quasi*-ordered pairs. We think of the quasi-ordered pair (3) as having the individual a as its first member. But this too is not strictly correct. Nor is it correct to say that the virtual class $x \ni (\text{—}x\text{—})$ is its second member. For strictly, when we say that (3) is a fact, we are speaking quasi-syntactically. We appear to be speaking of the individual a and the virtual class $x \ni (\text{—}x\text{—})$, but actually we are speaking of the expressions which L-designate them instead, as we see by a glance at the appropriate definiens. At the very best, we can say only that, in expressions of the form (4), we are speaking of a *quasi*-individual, a *quasi*-virtual class, and a *quasi*-virtual ordered couple of that quasi-individual with that quasi-virtual class!

Tenuous indeed then are facts, mere "trifles light as air." Nonetheless, whole expressions of the form (2) or (4) are given a very concrete significance in the definientia. Also the definientia are extremely simple, involving merely quantifiers over the expressions of L, defined notions of L-designation, and the semantical notions of truth and logical truth. Such minimal tools are required in almost all philosophical analysis anyhow. It is interesting that we can get along here with so little. It is as though God made logic, throwing in a little meta-logic for good measure, and all the rest is the work of the philosopher.

We need not claim the foregoing definition-schema as a final one. Possibly upon further reflection we shall wish to emend it, perhaps by adding further conditions. Perhaps some requirement of specificity, as suggested by Carnap, would be desirable. This we need not decide for the present. As in philosophy and science generally, so here—we posit principles and definitions hypothetically as a basis for further exploration, hoping that they will turn out to be both convenient and fruitful for the purposes at hand.

One direction in which emendation might immediately seem called for, however, hovers around the distinction between *basic* (or *observational*) and *theoretical* predicates.[26] We might wish "facts" to be expressible wholly in terms of the former. This requirement is well brought out in the main item under "fact" in the *O.E.D.*, namely, that a fact is "known by actual observation or authentic testimony, as opposed to what is merely inferred, or to a conjecture or fiction; a datum of experience, as distinguished from the conclusions that may be based upon it." Accordingly, we might require that expressions for facts contain only basic or observation predicates (and perhaps that the proper names, and the range of the individual variables, be restricted to only certain kinds of concrete objects, excluding theoretical entities). We commonly speak of the facts of experience, after all, but only in a picaresque sense of the "facts" of, say, quantum mechanics.

IX

Introducing facts in the way suggested embodies the features (i)–(x) put forward above as desiderata. We speak of facts only relative to the object-language under discussion, and within an appropriate semantical meta-language. Expressions for them are eliminable, and they are fictitious in an appropriate sense. The semantic notion of truth is needed in the definiens, as well as the L-semantical notions of analyticity and of being factually true. Facts are "logically intensional" entities and have an internal structure. Let us explore these last two points a little more fully.

We have not incorporated L-equivalence into the notion of a fact here, as Scheffler did. But, if desired, we can build it into the notion of the *identity* of facts. Two facts are to be identical if their respective quasi-factors are in appropriate senses L-equivalent. What are these appropriate senses, one for individuals, one for virtual classes?

Among the individual expressions are included primitive individual constants. For simplicity, let us assume that descriptions, phrases of the form "the so and so," are also introduced as primitives. Then sentences of the form

The so and so = the so and so

are logically true irrespective of how many so and so's there are. (Not so if descriptions are introduced by Russell's method via definition.) Likewise, where "$(\iota x.\text{—}x\text{—})$" reads "the one x such that —x—,"

$$(\iota x.\text{—}x\text{—}) = (\iota x.\ldots x\ldots)$$

is then also logically true provided

$$(\forall x)\ (\text{—}x\text{—} \equiv \ldots x\ldots)$$

is analytic. (Again, not so via Russell's method, where hypotheses of uniqueness are needed). With these preliminary explanations we can gain a clear meaning for the L-equivalence of individual expressions as follows.

The individual expressions now comprise primitively all the individual constants and all descriptive phrases. Then we may say that

a is **L-equivalent** to b

---

[26] See especially R. Carnap, "The Methodological Character of Theoretical Concepts" in *Minnesota Studies in the Philosophy of Science*, Vol. I (Minneapolis, University of Minnesota Press, 1956), pp. 38–76; C. G. Hempel, *Aspects of Scientific Explanation* (New York, The Free Press, 1965), esp. pp. 173 ff.; *The Philosophy of Rudolf Carnap*, ed. P. A. Schilpp (LaSalle, Ill., Open Court Publishing Co., 1963), *passim;* and the author's "On Theoretical Constructs and Ramsey Constants," forthcoming in *Philosophy of Science*, and "On Theoretical Entities," forthcoming in *Ratio*.

provided (i) a and b are the same primitive individual constant or the same description, or (ii) a = "(ιx.—x—)," and b = "(ιy. . .y. . .)," and "($\forall x$) (—x—$\equiv$. . .x. . .)" is L-true or analytic, where ". . .x. . ." is understood to differ from ". . .y. . ." only in containing free 'x' wherever and only where there are free 'y' in ". . .y. . . ." (The case where a is a primitive individual constant and b is a description need not be admitted, for then the individual constant could be introduced as an abbreviation for that description. Also for simplicity we assume that no two primitive individual constants designate the same individual.)

Similarly two virtual-class expressions a and b will be said to be L-equivalent provided a = "$x\ni$(—x—)," b = "$y\ni$(. . .y. . .)," and

($\forall x$) (—x—$\equiv$. . .x. . .)

is analytic (". . .x. . ." differing from ". . .y. . ." appropriately).

We can now go on to a schema for the identity of facts as follows.

<**a**,$x\ni$(—x—)> = <**b**,$y\ni$(. . .y. . .)>

abbreviates

<**a**,$x\ni$(—x—)>-is-a-fact, <**b**,$y\ni$(. . .y. . .)>-is-a-fact, a is L-equivalent to b, and c is L-equivalent to d,

where in place of '**a**' and '**b**' we put in either two primitive individual constants or two descriptions, in place of 'a' and 'b' we put in the structural-descriptive names respectively of those constants or descriptions, and in place of 'c' and 'd' we put in respectively the structural descriptions of "$x\ni$(—x—)," and "$y\ni$(. . .y. . .)."

Let us consider an example. Suppose "Scott" is a primitive individual constant and "the author of *Waverly*," and "the author of *Marmion*" are primitive descriptions (more properly "the one x such that x is an author of *Waverly*," and "the one x such that x is an author of *Marmion*"). We need not symbolize them more exactly. Then, to speak loosely,

(1) <Scott, $x\ni$(x lived at Abbotsford)>,
(2) <Scott, $x\ni$(x=the author of *Waverly*)>,

and so on, are facts. On the other hand

(3) <the author of *Waverly*, $x\ni$(x=the author of *Waverly*)>

is not a fact, because of the analyticity of "the author of *Waverly*=the author of *Waverly*."

Note that "Scott=the author of *Waverly*" is true but not analytically so. We cannot from this conclude that because (2) is a fact, (3) is then also. The reason is that "Scott" and "the author of *Waverly*" are used in (1)–(3) only quasi-syntactically, i.e., strictly they are not used at all, but only mentioned.

Similarly

<the author of *Waverly*, $x\ni$(x lived at Abbotsford)>

and

<the author of *Marmion*, $x\ni$ (x lived at Abbotsford)>

are distinct facts even though

the author of *Waverly*=the author of *Marmion*

is true (but not analytically so).

The internal structure of facts remains to be commented on. Of this both Russell and Wittgenstein made much. In failing to realize the intensional or quasi-syntactical nature of facts, they in effect confused use and mention. Hence they sought in the world for a structure that is to be found only in language. The structure of a fact can be read off from the virtual-class expression constituting its second member. And this may be of any complexity, atomic, disjunctive, implicative, negative. Are there then disjunctive, negative, etc., facts? Only as special instances of the general schema. This seems harmless enough, and there is no need for further perplexity on the subject.

Also we should observe that there is no need for introducing dyadic, triadic, etc., facts, all of these being included in the general schema. We see this by observing that relational sentences such as "a is the father of b" may also be handled as virtual-class sentences of the form "a is a member of $x\ni$ (x is the father of b)." Nonetheless, we may introduce dyadic, etc., facts explicitly, if desired. Thus

<a,$xy\ni$(—x—y—),**b**>

would be the kind of expression needed for dyadic facts. And so on.

X

Having found facts to be eliminable, fictitious, nominal entities, we might well wonder whether we should bother about them at all. Quine suggests something similar. "We were able," he writes, "to abjure sakes, measures, unactualized possibles, and facts without a pang, having satisfied ourselves that to admit them would serve no good purpose."[27] Abjuration for Quine consists of not recognizing items as values for variables. It need not consist in not recognizing a word or phrase as a *manière de parler*. Rather we should absorb or "paraphrase away" the various uses of the abjured words. This is precisely what we have tried to do here for "fact." We have concentrated on the use "so and so *is a fact*," as by far the most fundamental. In this we agree (apparently) with Austin. We can go on to "fact that"-talk in easy and natural ways. But for this a good deal needs then to be said about the so-called propositional attitudes, knowing, believing, and so on.[28]

We have a choice before us now. We can forswear all uses of "fact" with impunity, for they are all eliminable or paraphraseable anyhow. Or, alternatively, we can use it in various ways, subjecting each way to proper analysis, ultimately coming back hopefully to some such fundamental way as the one suggested. The view here is that it is not the aim of logical analysis to impoverish our everyday modes of speech. Let us rejoice in the richness of its vocabulary and the enormous range of its expressive power. But if it is a logical analysis that we want, nothing short of the deepest searching for fundamentals will suffice. For this we must have recourse almost always to a properly restricted and otherwise acceptable semantical metalanguage.

Even at the depth level, however, we have seen that there is no need of "fact" strictly. Instead we can always go back to "factually true sentence." Thus, in a sense, our whole enterprise is a thing of nought. If we have succeeded in reaching our goal, we have only shown that we did not have to start out in the first place. The holy grail is already to be found at one's doorsteps. But perhaps in having made the search, we shall come to value "factually true" the more highly and to use it the more carefully.

New York University

Received June 9, 1966

---

[27] *Word and Object, op. cit.*, p. 248.

[28] See especially *Belief, Intension, and Ontology, op. cit.*

American Philosophical Quarterly
Volume 4, Number 4, October 1967

III. WHY I KNOW SO MUCH MORE THAN YOU DO

WILLIAM W. ROZEBOOM

WHAT does it mean to say that person X "knows" that p? With a unanimity remarkable for philosophers, it is generally agreed that for this to be true, it must obtain that

(a) p is the case,
(b) X believes that p, and
(c) X is justified in believing p.

Whether conditions (a)–(c) jointly suffice for X to know that p, however, has recently been disputed by several writers.[1] I shall attempt to show that these doubts are unfounded, and that the *justified true belief* analysis of knowledge (hereafter referred to as the "JTB thesis") is indeed adequate. But I shall then proceed to agitate related perplexities about the concept of "knowledge" and conclude with a possibly heretical suggestion about its continued usefulness for technical epistemology.

I

Estranged from his wife and beset by financial troubles, John Duosmith has become conspicuously despondent. Today, a man's body is found in Duosmith's hotel room with Duosmith's revolver in its hand, a bullet therefrom in its head, and a suicide note signed by Duosmith on the table. Mrs. Duosmith identifies the body as that of her husband, pointing out his characteristic birthmark, the private details of their recent quarrel cited in the note, and so on for many other items which make it overwhelmingly evident to Mrs. Duosmith that the corpse of John Duosmith lies before her, and hence that her husband is dead.

And John Duosmith is indeed dead. But what has happened is this: Last night, Duosmith received a secret visit from his identical twin brother Jim, a petty criminal whose existence John had concealed from his wife and who now begged John to hide him from certain colleagues seeking retribution for something Jim had done. Seeing a chance to make a new life for himself, John shot his brother and arranged the scene to appear as though he, John, had killed himself. But as John left the hotel, he was spotted by Jim's pursuers who, mistaking him for his twin, promptly executed their plans for Jim's demise. So John Duosmith is dead while his wife, for the best of reasons, also believes this to be so.

But does Mrs. Duosmith *know* that her husband is dead? Mr. Gettier and others[2] say "No," and my own linguistic intuition agrees with this judgment. The force of this and similar examples cited by Gettier *et seq.* is drawn from the principle that true beliefs grounded upon false premisses do not count as knowledge, no matter how reasonable those premisses may themselves be under the circumstances. That is,

(A) If person X believes p—justifiably—only because he believes q, while he justifiably believes q on the basis of evidence e, then q

---

[1] See Gettier (8), Clark (7), Sosa (17), Saunders and Champawat (16), Lehrer (12), and Harman (9). The justified-true-belief view of knowledge has been well stated by Ayer (3, ch. 1), Chisholm (5, ch. 1), and Woozley (18, ch. 8, though Woozley rashly puts the justification condition as having *evidence* for what one knows), and recently defended in one respect or another by Arner (2), Saunders (15), and Harrison (10). While condition (b) has occasionally been disputed on the rather foolish ground that "believes" is sometimes understood to imply "does not feel sure of," the necessity of (a) and (b) for knowing that p is essentially noncontroversial. The status of (c), however, is more problematic. Armstrong (1, p. 120), Malcolm (13, p. 225ff.), and Sosa (17) contend from an overly narrow equating of "justification" with "evidence" that justification is not always requisite to knowing. In contrast, Gettier (8), Clark (7), Lehrer (12), and Harman (9) accept the necessity of (c) for knowing p but deny its sufficiency given (a, b), while Saunders and Champawat (16) question the possibility of finding *any* set of conditions which are necessary and sufficient for all instances of "knowledge."

Though not strictly addressed to the analysis of "knowing," recent discussion by Chisholm (6), Brown (4), and Saunders (14) concerning self-justifying beliefs, and Hintikka's (11) widely acclaimed exploration of the modal logic of knowledge and belief are also background context for the present work.

[2] See n.1.

> as well as p must be the case if X's belief in p is to qualify as "knowledge."

Consequently, if p is true while q is false in such a case, as apparently illustrated by the Duosmith episode and Gettier's examples, it follows that a person may justifiably believe true proposition p and still not *know* that p. Now in fact, these examples do *not* show this, nor can the JTB thesis ever be threatened by principle (A). But before I point out why this is so, it is best to undermine confidence that linguistic intuition can be trusted to provide a sound interpretation of cases like these.

II

It is Sunday afternoon, and Mrs. Jones is on her way to borrow an egg from her neighbor, Mrs. Togethersmith. She fears that she may be too late, however, for she is aware that every Sunday afternoon for the past several years, the Togethersmith family—Mr. and Mrs. Togethersmith and their two children—has gone for a drive in the country. As she steps outside, Mrs. Jones sees the Togethersmith car departing with Mr. Togethersmith at the wheel, and thinks to herself, "Pity, there she goes." Mrs. Jones believes that Mrs. Togethersmith is driving away because, for excellent reasons, she believes that the entire Togethersmith family is in the departing car. But in fact, while Mrs. Togethersmith, her husband, and one of their children are indeed in the departing car, the other Togethersmith child is on this one occasion attending a friend's birthday party. Insomuch as it is not true that the entire Togethersmith family is driving away, is Mrs. Jones's justified true belief, that Mrs. Togethersmith is driving away, *knowledge?* Principle (A) appears to deny this, since Mrs. Jones arrived at her true belief about Mrs. Togethersmith by means of a justified but false belief about the whereabouts of the entire Togethersmith family. But the falsehood here seems so *irrelevant.* For "The entire Togethersmith family is driving away" is equivalent to the conjunction "Mrs. Togethersmith is driving away, Mr. Togethersmith is driving away, and all the Togethersmith children are driving away," the components of which are supported by Mrs. Jones's evidence just as well separately as conjoined—in fact, it is difficult to say whether Mrs. Jones's belief about Mrs. Togethersmith derives from her belief about the Togethersmith family as a whole, or is a part-cause of it. In any event, linguistic intuition is disposed to deny that the absence of one Togethersmith child from the departing family car disqualifies Mrs. Jones's justified true belief about Mrs. Togethersmith's departure as an instance of knowledge.

But if so, what about Mr. Jones, who, wanting a 12-inch board, measures one with his new tape rule, obtains a 10-inch reading, and concludes "That's too small," when the tape rule is defective and this board is really 11 inches? (We assume that Jones has had much past experience with tape rules, all of which amply warrants his trust in the present reading.) Mr. Jones's justified belief, that this board is under 12 inches, is true even though it is derived from his false justified belief that this board is 10 inches. Since intuitively this case is no different in kind from Mrs. Duosmith's belief about her dead husband, we should deny that Jones *knows* this board is less than 12 inches. Yet "This board is 10 inches" is equivalent to the conjunction "This board is under 12 inches, this board is at least 10 inches, and this board is not between 10 and 12 inches," only the last component of which is false while its first component requires for its justification only a proper part of the evidence which supports Jones's belief in the whole conjunction. So by formal parallel, it might also seem that Mr. Jones's conclusion that this board is undersize should not be epistemically inferior to Mrs. Jones's belief in Mrs. Togethersmith's departure.

More intensive analysis of these two cases would show not so much that one or both violate principle (A) as that when a person's justified true belief p is accompanied by a justified false belief q, it may well prove troublesome to decide whether or not his belief in p is related to his belief in q in such fashion that the falsity of the latter should disqualify the former as knowledge. That this is in general a failure of conception, not just an insufficiency of data concerning the believer's detailed reasoning, is shown by the following more sophisticated example.

Dr. Pillsmith, a competent practitioner of medicine, is well aware that

(1) Among persons who have not been vaccinated against Hypofluvia, 999,999 out of a million who show symptoms S are afflicted with this disease,

(2) Among vaccinated persons showing symptoms S, only one in ten is afflicted with Hypofluvia,

(3) Only one person in a million showing symptoms S has been vaccinated against Hypofluvia,

and that consequently,[3]

(4) More than 999,998 persons in a million who show symptoms S are afflicted with Hypofluvia.

Attempting to diagnose the condition of his latest patient, Dr. Pillsmith observes that

(5) Philip Blotely shows symptoms S,

and, lacking further information about Blotely's medical history, infers unhesitatingly from (1)–(5) both that

(6) Philip Blotely has not been vaccinated against Hypofluvia

and that

(7) Philip Blotely has Hypofluvia.

Now it so happens that Blotely was, in fact, vaccinated against Hypofluvia, but has contracted it just the same. So Pillsmith's diagnosis (7) is both justified and true—but is it knowledge? At first it might seem that the falsity of (6) thwarts this, for were Pillsmith to surmise the truth of Blotely's vaccination, his knowledge of (2) and (5) would prevent him from justifiedly accepting (7). Yet (6) is at the same time irrelevant to the diagnosis in that Pillsmith can get to (7) from (5) and (1)–(3) by way of (4) without ever considering whether or not Blotely has been vaccinated. And if Pillsmith does make his diagnosis in this way, must (6) still be true if Pillsmith's justified belief, that Blotely has Hypofluvia, is to count as knowledge? Surely not, since falsehood (6) takes no part in the inference. Yet we can also argue that justification of (7) by (5) and (1)–(3) via (4) implicitly presupposes the truth of (6), for derivation of (4) from (1)–(3) argues in effect: Given any person with symptoms S, either he has or has not been vaccinated against Hypofluvia. If he hasn't, it is virtually certain that he has Hypofluvia; otherwise, Hypofluvia is counterindicated, but this is too unlikely a possibility to be considered seriously.

It seems to me that the intuition which was so sure in Sect. I that Mrs. Duosmith doesn't really know her husband is dead is quite at a loss to say whether or not Dr. Pillsmith knows that Blotely has Hypofluvia. There is quicksand underfoot here, and we must not too hastily presume that Duosmith-type examples, which apparently refute the JTB analysis of knowledge by way of principle (A), are all that they seem to be.

III

It is now time to make explicit an important technical detail which is usually slighted in philosophical discussions of knowledge. This is that the judgmental attitudes in which a proposition can be held are not just belief and disbelief, or belief, disbelief, and uncertainty, but a whole spectrum of credibilities spanning many shades of uncertain belief and doubt. Consequently, assertion that knowing presupposes believing is specious unless it is made clear just how strong a belief is so required. Once this question of degree is raised, however, we can easily see from the absurdity of "He knows that p but isn't entirely sure of it,"[4] or "I know that p but have some doubts about whether it is really so," that only maximal belief is acceptable for knowledge. Hence condition (b) of the JTB thesis must be explicated as "X feels completely sure of p" or "X believes p absolutely," while similarly, (c) must be read as "X is justified in believing p absolutely," "X has a right to have not the slightest doubt about p," or the like. (Also, since it may be argued that if X is justified in believing p in degree d then X is also justified to almost the same extent in believing p to a degree which is almost d, "X is justified in believing p absolutely" should be further explicated as "X has more justification for believing p absolutely than for feeling any doubt about p.")

When is absolute belief justified? While a convincing general answer is not easy to come by (see Sect. IV), a necessary condition for evidential justification is surely the following:

(B) If person X feels completely sure of p on the basis of evidence e, then X's belief in p is justified only if e necessitates p.

That is, X is not justified in feeling certain of p in virtue of his awareness of e unless p *is* certain, given

---

[3] Since for any three attributes A, B, and C, $\Pr(A/B) = \Pr(C/B) \times \Pr(A/BC) + \Pr(\bar{C}/B) \times \Pr(A/B\bar{C}) \geq [1 - \Pr(C/B)] \times \Pr(A/B\bar{C})$.

[4] At first thought, this might seem to make sense as a variant of "He really knows that p but can't bring himself to admit it." But *not admitting to belief in* is not at all the same as *having some doubts about*, and "He knows that p but won't admit it" implies not "He is not sure of p and won't admit that p," but either "He doesn't really have any doubt about p but can't bring himself to say so," or "He still isn't really convinced of p even though he has overwhelming evidence for it."

e. This is entirely compatible with admitting that *X* may be justified in feeling *almost* certain of *p* on grounds *e* if *p* is extremely likely, given *e*. It only denies that it is rational for *X* to close his mind completely to the possibility of not-*p* even though *e*, so long as this possibility does in fact exist. Moreover, while (*B*) does not specify what sense of necessity—logical, causal, or whatever—is required, it is in any case analytically true that

> (*C*) If *e* while not *q*, then it is not the case that *e* necessitates *q*.

Finally, I think it will be agreed that in general,

> (*D*) If person *X* feels completely sure of *p* only because he feels completely sure of *q*, but his belief in *q* is not justified, then neither is his belief in *p* justified.

(I would hold that (*D*) is always the case, but there is room for argument on this point when *q* is "basic" for X—see Sect. IV—and it is not essential here that (*D*) be completely universal.)

Let us now reconsider principle (*A*), which envisions a person's believing *p* on the basis of his belief *q*, and *q* on the basis of evidence *e*. For this to pose any threat to the JTB analysis of knowledge, the degree of belief at issue must be *absolute* belief. But it follows from (*B*) and (*C*) that *X* can justifiedly believe *q*—absolutely—on the basis of evidence *e* only if *q* is the case; so stipulation of *q*'s truth in the final clause of (*A*) is in this case otiose. That is, if *X* feels completely sure of *p* only because he is convinced of *q*, and the latter only because he is aware of *e*, then to hypothesize that *q* is false is also, by (*B*) and (*C*), to presuppose that *X*'s absolute belief in *q* is unjustified, and hence by (*D*) that neither is he justified in feeling completely sure of *p*. Thus in our Duosmith example (and similarly for Gettier's cases), while Mrs. Duosmith had excellent reason to feel *virtually* certain that her husband was dead, the bare fact that, overwhelming evidence to the contrary notwithstanding, the body before her was not that of her husband shows that this evidence did not warrant her having no doubt whatsoever that her husband was dead. Likewise, for our more problematic examples in Sect. II, we can say without hesitation that Mrs. Jones, Mr. Jones, and Dr. Pillsmith did not *know* that Mrs. Togethersmith was departing, that this board was less than 12 inches, and that Blotely had Hypofluvia, respectively, because while the evidential bases for these conclusions made them extremely likely, a vestige of uncertainty still remained. In short, if the "belief" cited in principle (*A*) is allowed to include degrees of belief weaker than complete conviction, the truth of (*A*) resides in that the falsity of *q* is symptomatic that the degree of *p*-belief justified by *e* is less than absolute.

IV

But while the argument from principle (*A*) thus fails to impeach the JTB thesis, the claim that justification is always prerequisite to knowledge is far from unproblematic. Most conspicuously troublesome is that if *X*'s knowing *p* requires there to exist evidence *e* such that *X*'s belief in *p* is justified by *X*'s awareness (i.e., knowledge) of *e*, then *X*'s belief in *e* also requires such justification and we are off on a regress. In particular, the justification requirement might seem to exclude the possibility of perceptual knowledge and self-awareness where *X*'s belief in *p* is not inferred from other beliefs but is aroused directly by sensory stimulation or given introspectively. Moreover, as will be seen, the demand for justification undergoes a remarkable transformation when we turn from other-person to first-person knowledge, while the conditions of justification even for the inferred beliefs of others are not so straightforward as they might at first appear.

Actually, no puzzle of "justification" can even discredit the JTB thesis, for the simple reason that whatever is needed for *X* to know that *p*, if *X* does know *p* then he is certainly justified in believing *p*. (Witness, e.g., the absurdity of "*X* knows that *p*, but he has no right to believe it so strongly.") So we can maintain that *X* knows that *p* iff *p* is a justified true belief of *X*'s without concern for how murky the concept of justification may itself be. I submit, therefore, that the greatest philosophic challenge which issues from the JTB position is not to settle whether this view is entirely correct (though the import of my argument in Sect. III is that we have no good reason to doubt this), but to determine what it is about *X*'s knowing *p* that accredits *X*'s *p*-belief as "justified." The intent of this section is to rough out a tentative solution to this problem, which is more untidily complex than heretofore recognized, in full expectation that many more exchanges will be needed to round out the present survey in convincing detail.

Let a person's belief in *p* be described as "basic" if he does not believe *p* as a result of his believing something else. (For example, when *X* perceives that p, his belief in p is simply for him a *given* which becomes a basis for inference but is not derived from anything else he knows.) Then X's *basic convictions* are beliefs of which X feels completely sure without having inferrred them from evidence. Unless knowledge can be inferred from beliefs which do not themselves qualify as knowledge (a counterintuitive possibility which I shall not discuss), the regression argument shows that if X knows anything at all, he must also have basic knowledge, i.e. justified true basic convictions, and our first task in this section is to find some acceptable sense in which a basic conviction may be said to be justified. Two alternatives present themselves: either (1) basic convictions are self-justifying, or (2) some basic convictions have nonevidential justification.

In support of (1), it might be argued that the justification of X's belief in p consists in p's bearing a certain relation J to some set B of X's true basic convictions. If J is such that each $b_i \in B$ is also related to B in manner J, then all basic convictions in B are also by definition justified. For example, if "X is justified in believing p" were to be analyzed as "p is logically entailed by X's true basic convictions," then X's basic convictions are justified by the reflexivity of entailment. But this approach leaves much to be desired. For one, an intuitively acceptable J-relation with the needed formal properties is not easy to come by. (It is simple to argue, e.g., that logical entailment does not in itself confer evidential justification.) Further, how is the set B to be circumscribed? Does veracity suffice for a basic conviction to belong to B—i.e., for it to be justified? If so, we should have to grant that a dogmatic thinker who habitually works himself into a state of absolutely closed judgment on controversial issues without considering any of the relevant evidence is justified in holding any such belief which by chance happens to be true. But if we hold that a basic conviction can be true without necessarily being justified, whatever else is needed for the latter constitutes a nonevidential source of justification and thus carries us into alternative (2).

There are many intriguing thought experiments by which our intuitions about nonevidential belief-warrants can be bared, starting with increasingly bizarre or futuristic ways (e.g., electrical stimulation of the retina) in which sensory input might elicit true perceptual beliefs; but here it will suffice to consider just one which cuts directly to the heart of the matter. Suppose that Tom Seersmith claims to be able to foretell the outcomes of horse races. Upon investigation, we learn that when Seersmith thinks about a forthcoming race, he is often overwhelmed, quite without any reason for it, with a feeling of complete certainty that a certain horse will be the winner. Before the last Kentucky Derby, Seersmith felt sure that it would be won a by horse named Fleetfoot, and as it turned out his prediction was correct. Did Seersmith *know* that Fleetfoot was going to win, and if so, in what sense was his belief justified?

Whatever the personal peculiarity which endows Seersmith with convictions about forthcoming horse races, if his prognostication record has shown only chance accuracy in the past, we would be loath to say that Seersmith either knew or was justified in believing that Fleetfoot would win the Derby even though his belief in this instance happened to be true. Even if Seersmith's previous race predictions have usually been correct, with a hit rate high enough to convince us that there is something extraordinary about this man, we would still deny that he was justified in feeling absolutely sure that Fleetfoot would win so long as his predictions are not infallible. But suppose we discover that Seersmith's horse race prognoses *are* infallible —i.e., we become convinced that whenever Seersmith feels sure that race r will be won by horse h, it is absolutely certain that h will win r. Then surely we would be forced to admit that Seersmith knew that Fleetfoot would win the Derby, even though how he knew would baffle us. (As I follow Seersmith through prediction after prediction and see that he is *never* wrong, I find myself saying, "I can't understand it, but somehow, he *knows*!") The justification—nonevidential—for Seersmith's belief is simply that since generalization

> $(\forall h)(\forall r)$(Tom Steersmith believes that horse h will win race $r \supset h$ wins r)

is a nomological principle of our world, Seersmith's basic conviction that Fleetfoot would win the Derby was true not by mere happenstance but of nomic necessity. For if Seersmith's horse race precognitions *cannot* be wrong, what better justification could there possibly be for his having them?

Now let's change the case slightly. Tom Seersmith's brother Dick also feels occasional convictions about the outcomes of forthcoming horse races, but unlike his brother, Dick's percentage of correct anticipations is substantially less than perfect. Careful research discloses, however, that Dick's accuracy depends critically upon the horse's name. Whenever Dick feels sure that horse *h* will win race *r*, he is never mistaken so long as *h*'s name contains exactly two syllables, but when the predicted winner's name is shorter or longer than this, Dick's precognitive effectiveness is somehow impaired. Dick, too, felt sure that Fleetfoot would win the Derby—but did he *know* this? In principle, Dick's case is exactly like that of brother Tom, since by natural law his preconvictions about bisyllabically designated horse-race victors cannot err. Yet intuition is more hesitant here, for at times Dick also feels certain that *h* will win *r* when he should not, namely, when '*h*' is not bisyllabic. And if Dick is unaware that the anticipated winner's name makes a difference for the reliability of his forecast, we might question whether he is entitled to feel such perfect confidence in his precognition even when, unknown to him, it is in fact nomologically infallible. Thus it might be denied that Dick *knew* of Fleetfoot's forthcoming victory if he did not recognize that this particular belief had stronger truth-credentials than his average forecast. But this line of argument is unsound. If Dick were aware of his general precognitive fallibility, it would indeed seem reasonable for him to have inferential meta-doubts about his belief in Fleetfoot's victory—if he had, in fact, had any. But to make a person's knowing *p* contingent upon his knowing that he knows *p* would precipitate an intolerable regress, nor should a person's true belief in *p* be disqualified as knowledge merely by his having additional erroneous convictions as well. (To hold that a person can know nothing if he ever believes falsely seems extreme to the point of absurdity, though as will be seen in Sect. V it contains an important grain of truth.) If Dick's conviction that Fleetfoot would win the Derby was truly basic for him, uninfluenced by any meta-beliefs concerning his general prognostic proficiency, then it was for him no less an instance of *knowing* than it was for brother Tom, and was justified on the very same grounds, namely, that insomuch as Dick felt certain Fleetfoot would win, it *was* certain that Fleetfoot would win.

And now one more twist. Suppose that Harry is still another Seersmith whose case is like Dick's except that Harry's precognitions of form "horse *h* will win race *r*" are always (nomologically) correct when and only when *r* is run on a dry track. Insomuch as Fleetfoot won the Derby on a dry track, was Harry's conviction that Fleetfoot would win an instance of knowledge? Harry's belief, too, was infallible in that there exists a nomic principle in virtue of which, given that Harry felt sure that Fleetfoot would win and that the track was dry, it was certain that Fleetfoot would win. But Harry differs from Dick in that Dick's infallible precognitions are intrinsically identifiable as such—i.e., whether or not a forecast by Dick falls under the law which vouchsafes its accuracy is revealed by its syllabic composition—whereas the reliability of Harry's forecast cannot be determined without additional information which does not generally become available until race time. And since Harry thus cannot discriminate his infallible precognitions from those which are not (assuming that he does not also have advance knowledge of the track conditions), it might be argued that he had no right to feel so sure that Fleetfoot would win the Derby. But this doubt is only a refinement of the one we have already rejected in Dick's case. If it is not necessary for Harry to know that he knows *p*, or to know that he can trust his belief *p* absolutely, in order for him to know *p* or for his *p*-belief to be justified, then neither can we reasonably hold that his being *able* to acquire this meta-knowledge is requisite for the latter. (How could such an ability possibly be germane except by way of the knowledge which, with its help, Harry does acquire?) Consequently, so long as a person's basic believing of *p* belongs to a class whose members are nomologically infallible, its epistemological status should not depend upon whether or not this class is defined by properties inherent in the belief itself. (It is relevant to this point that knowing what *we* know about Harry, we could win a pretty penny at the races by noting Harry's prediction and waiting to see the track conditions before deciding whether to bet.)

The principle which appears to govern our Seersmith examples (though intuition speaks only with a subdued and halting voice here) is that a basic conviction counts as "knowledge," and is by the same token justified, if and only if it not merely *is* true, as could occur by chance, but is infallibly true by virtue of its being of a kind (perhaps defined in part by relational attributes) whose accuracy is guaranteed by natural law. This is at best an uneasy conclusion, however, for without further qualification it trembles on the brink of triviality: If V is the attribute of veridicality, then any belief of type V—i.e., any which happens to be true—is also infallibly true vis-à-vis type V and hence qualifies as knowledge unless "natural law" is defined to exclude generalizations which are logical truths. (That we would *like* to make some such exclusion is, I think, intuitively evident, but how to accomplish it effectively is another question.) Moreover, this criterion applies only to the basic convictions of other persons, for *my* beliefs, basic or otherwise, are justified by standards rather different from this.

Suppose that I set out to determine which of us, you or I, knows the more. I start by listing all propositions which you believe (or more precisely, all which I believe that you believe) and then prune this list by deleting everything on it which, in my judgment, is either false or is unwarranted for you. To display what *I* know, however, (or rather, what I believe that I know), I list all the propositions believed by me—*and stop*. For while I am perfectly willing to admit that there are facts of which I have no knowledge, I do not believe, nor can I bring myself to believe, any specific proposition of form "*p*, but I don't know that *p*." This is closely related to the oddity of "*p*, but I don't believe it," but has a significance the latter does not. I reject "*p*, but I don't believe it" because my believing this would entail my believing *p* and hence falsify the conjunction as a whole—i.e., it is impossible for me to believe a true proposition of this form. I could, however, truly believe "*p*, but I don't know it" were *p* to be the case while I believed but did not know that *p*.[5] Hence my refusal to admit, when I am convinced of *p*, that I may not know that *p*, has the force of maintaining that in *my* case, believing truly suffices for knowing. (And yet, when I reflect upon why I don't consider all *your* true beliefs to be knowledge, I am also willing to admit in general terms that some of mine may not be knowledge, either. This is an apparent inconsistency which will be resolved in Sect. V.)

This first-person/other-person difference in knowledge criteria is present even when, from your vantage point, my true belief is amply justified. Suppose that I believe mathematical theorem T because I have just discovered a convincing proof of it, and when I inform you that I know T, you ask me what grounds I have for thinking that I know this. The only reply which seems relevant is for me to recapitulate the steps by which I deduced T—except that when I do this, you quite properly point out that what I have given you is grounds for believing T whereas what you asked for is grounds for believing that I know T. My supplying of proof for T *demonstrates* to you that my belief in T is justified, but *asserts* nothing which implies this. On the other hand, if I try to give you an account of how it is that I know T, it will go something like: "Well, I know that L_1 and L_2 are logically true, so when I see that T is an immediate consequence of L_1 and L_2, this makes me aware that T is also the case." What I am telling you is that my awareness of certain facts, namely, that L_1 and L_2 are logically true and jointly entail T, is a *cause* of my knowing T. But from your perspective, my awareness of this evidence *provides* (noncausally, by fulfilling an existence requirement) the justification which is an analytic component of my knowing T. That is, in more general terms, when I have come to know *p* by a valid line of reasoning from unimpeachable premisses, the inferential procedure which to you is a *condition* on my knowing *p* is to me only the *occasion* for this. What justifies *my* believing *p* is simply *p*'s being the case; for once I have convinced myself that *p*, I require nothing else to consider it right that I feel sure of *p*—what more could possibly be relevant? In fact, it is nomically impossible for me to accept "*p*, but I have no right to believe it so absolutely," for my believing the second component of this suffices to create some doubt in me about *p*. (This is why I am unable—nomically, not logically—to believe "*p*, but I don't know it.") So I reason about *p* not to transmute my base belief in *p* into golden knowledge, but to decide whether *p* is the case; whereupon, having convinced myself that it is, I rebuff all challenges to my conviction's epistemic credenials by the argument "*p*; therefore it is reasonable for me to believe that *p*."

What are my criteria for the justification of *your* beliefs? We have already explored the evaluation of your basic convictions, but it remains to see what is needed for your inferred beliefs to be reasonable. Ordinarily, the degree of *p*-belief which I consider evidence *e* to warrant in you is simply the confidence in *p* which I sense is aroused in me by conviction that *e*. However, closer analysis shows this not to be definitive. Let *p* and *q* be two propositions such that *q* logically entails *p*. If, somehow, I know that you know *q*, what will persuade me that you have evidential justification for believing *p*? For myself, merely being cognizant of q does not in itself convince me of p (I am, after all, often unsure of consequences of my beliefs when I have not discerned that they are such consequences), so for your q-awareness to warrant your p-belief, I would normally require you to know not only that q, but that q entails p as well. But is q & $(q \vdash p)$ then *sufficient* evidence to justify your belief in p? If, either thoughtlessly or with good reason, I presume that our minds work alike, I will agree that it is; for belief in the former would cause me to believe the latter as well. But suppose I have learned that you just don't seem to grasp the significance of logical relationships; specifically, that you often accept propositions of form "α, and α entails β" while simultaneously doubting or even disbelieving β. Then I would no longer consider your belief in q & $(q \vdash p)$ to be evidential justification for your believing p, for I could not regard the former as the *source* of the latter, nor would I have reason to regard your p-belief as knowledge insomuch as your veridicality in this instance may simply be the chance success of irrational thought. Conversely suppose I discover instead that your thinking is superrational in that for any two propositions α and β such that α entails β, and you believe α, your mere thinking of β when you believe α suffices for you to be convinced of β as well—i.e., you believe all logical consequences of your beliefs, so long as they come to mind at all, whether or not you are also aware of the entailment relations among these propositions. In this case, I must concede that your knowledge of q is fully adequate in itself to justify, and hence to dignify as knowledge, your belief in p. For if you accept all logical consequences of your convictions, rather than just a proper subclass of these constrained by your logistical perceptiveness, on what grounds can I hold your belief processes to be epistemically defective?

---

[5] While Hintikka (11) has proposed essentially the same analysis of "*p*, but I don't believe it" as offered here, his modal system allows that a person *can* defensibly claim to believe "*p*, but I don't know it." However, Hintikka's intuitive apologia for this (11, p. 83) construes it to involve a degree of *p*-belief less than perfect conviction, and the fact that his system does not recognize the unacceptability of "I don't know that *p* even though I am absolutely sure of it" would seem to reveal a lacuna in its axioms.

The preceding arguments show, first impressions to the contrary notwithstanding, that my other-person epistemology has exactly the same justificational standards for inferred beliefs as it does for basic ones. In both cases, the critical determinant is not whether the belief in question is "reasonable" in accord with some impersonal normative ideal, but whether it has arisen in circumstances which guarantee its accuracy. (This applies also to beliefs with derivational status intermediate between basic and inferred, thus bypassing the problems these would generate were the criteria of justification qualitatively different for the two extremes. For example, suppose that for a certain pair of attributes P and Q, whenever you perceive that $P(a)$ for an object a you find yourself also convinced that $Q(a)$. If I know that in our world P nomically implies Q, I must count your belief in $Q(a)$ as knowledge, justified by your awareness of $P(a)$, even if you are not consciously aware that $(\forall x)[P(x) \supset Q(x)]$.[6]) Harmony between first-person and other-person justification, however, is more elusive. At one level, these show a common pattern: Whenever I consider my/your conviction in p to be justified, I presume there to exist an argument of form "S; hence my/your belief in p is necessarily correct," where S is some state of reality.[7] But when it is my belief which is at issue, S is p itself; whereas I want the S which warrants your p-belief to do so nomically rather than logically. That I should adopt so flagrant an epistemological double standard is not a conclusion which I find esthetically pleasing, but if there is a deeper unity here I have yet to find it.

V

However lacking in clarity (and perhaps consistency) the epistemic concept of "justification" may be, it nonetheless appears that knowing p analytically requires not only that the knower feel completely certain of p, but also that there be some sense in which, considering the circumstances, it *is* completely certain that p. But since I reject the argument "p; therefore you are justified in feeling sure of p," and also doubt that our *de facto* world contains any nomic regularities perfect enough to vouchsafe any belief beyond all possibility of error, I do not think that you strictly know anything at all. Whereas in my own case, I have too much faith in my own fallibility to feel *absolutely* sure of anything, even if some of my perceptual beliefs fall short of this only negligibly. My admission (with high but not complete conviction) that probably not all of my beliefs are entirely correct, plus inability to meta-distinguish those which are true from those which are not, causally prevents me from ever entirely achieving an absolute extremity of belief, while my professional skills as scientist and philosopher enable me to find genuine even if minuscule chinks of uncertainty in any proposition I examine, even those which arise perceptually or feel analytically true. So technically speaking, I know nothing either.[8]

---

[6] I say "not consciously aware" rather than simply "unaware" to suggest the glide from inference episodes of the most paradigmatically rational sort down to believings which are patterned *as though* they were accompanied by additional supportive knowledge which the believer does not, in fact, have in any conceptualized form.

[7] The scope of "necessary" here is of course $\mathcal{N}(S \supset$ my-or-your p-belief is correct), not $S \supset \mathcal{N}$ (my-or-your p-belief is correct).

In short, my conception of "knowledge"—and presumably yours as well—is so impossibly idealized that no real-life belief episode ever satisfies it. Whenever you or I assert, as we often do, "I am aware that...," "He knows that...," etc., we are uttering falsehoods which would come closer to the truth if revised as "I approximate awareness that...," "He almost knows that...," or the like. The paradigm-case rejoinder, that what we *mean* by "know" is defined by these ordinary-life applications, no more shows that this usage is literally correct than the everyday paradigmatic ascriptions of "spherical" to roundish objects of irregular curvature demonstrates that a thing's surface does not really have to be a constant distance from its center in order for it to be literally a sphere. On matters philosophical as well as scientific, ordinary language teems with simplistic presuppositions and coarse-grained, uncritical categories which do slovenly justice to reality; and intellectual maturiry —represented most illustriously by technical science but by no means restricted thereto—consists first of all in learning to relinquish these cognitive crudities for a more sophisticated grasp of complexity and precise detail. It is all very well to recognize that the conceptual fluency of idealized approximations is often more convenient for everyday affairs than is the encumbrance of needless exactitude, but it is folly to construe the success of this practical usage as a sign that what is so asserted is precisely correct, or to begrudge its abandonment when, like outgrown clothing, its inaccuracies begin to chafe. In particular, there is no more reason for us to agonize philosophically over the esoterics of everyday knowledge-talk—e.g., why justified true beliefs at practical levels of assurance should sometimes be called knowledge and sometimes not—than for geometricians to puzzle over why some common-sense spheres have a larger cubed-surface-to-squared-volume ratio than do others.

To conclude, then, I propose that the subject of "knowledge" is no longer of serious philosophical concern for the simple reason that this concept is far too primitive for the needs of technical epistemology. No harm will be done, I suppose, by retaining a special name for true beliefs at the theoretical limit of absolute conviction and perfect infallibility so long as we appreciate that this ideal is never instantiated, but such sentimentality must not be allowed to impede development of conceptual resources for mastering the panorama of partial certainties which are more literally relevant to the real world. So far, however, the normative theory of practical belief has scarcely advanced beyond surmise that the structure of propositional credibilities is isomorphic to the probability calculus, and has not even begun to think technically about such vital subtleties as the ramifications of uncertainty in basic beliefs, reciprocal nondemonstrative supports among partially confirmed propositions, the credibility interplay between beliefs and meta-beliefs, and the like. With problems of "How strongly should X believe p?" lying dark and unfathomed before us, we stand to profit from continued epistemological preoccupation with the nature of "knowledge" to just about the same extent as would psychology from a return to study of the "soul."

University of Alberta

Received August 14, 1966

REFERENCES

1. ARMSTRONG, D. M. *Perception and the Physical World.* London, Routledge & Kegan Paul, 1961.
2. ARNER, D. "On Knowing." *The Philosophical Review*, vol. 68 (1959), pp. 84–92.
3. AYER, A. J. *The Problem of Knowledge.* London, Macmillan, 1956.
4. BROWN, R. "Self-Justifying Statements." *The Journal of Philosophy*, vol. 62 (1965), pp. 145–150.
5. CHISHOLM, R. M. *Perceiving: A Philosophical Study.* Ithaca, Cornell University Press, 1957.
6. CHISHOLM, R. "Theory of Knowledge." In: *Philosophy*, R. M. Chisholm, *et al.* Englewood Cliffs, Prentice-Hall, 1964.

---

[8] Hence the title of this paper is something of a misnomer. Although my knowledge criteria are enormously more liberal for me than for you, their extension is in both cases the null class.

7. CLARK, M. "Knowledge and Grounds: a Comment on Mr. Gettier's Paper." *Analysis*, vol. 24 (1963), pp. 46–48.

8. GETTIER, E. L. "Is Justified True Belief Knowledge?" *Analysis*, vol. 23 (1963), pp. 121–123.

9. HARMAN, G. H. "Lehrer on Knowledge." *The Journal of Philosophy*, vol. 63 (1966), pp. 241–246.

10. HARRISON, J. "Does Knowing Imply Believing?" *The Philosophical Quarterly*, vol. 15 (1963), pp. 322–332.

11. HINTIKKA, J. *Knowledge and Belief*. Ithaca, Cornell University Press, 1962.

12. LEHRER, K. "Knowledge, Truth and Evidence." *Analysis*, vol. 25 (1965), pp. 168–175.

13. MALCOLM, N. *Knowledge and Certainty*. Englewood Cliffs, Prentice-Hall, 1963.

14. SAUNDERS, J. T. "Beliefs Which are Grounds for Themselves." *Philosophical Studies*, vol. 16 (1965), pp. 88–90.

15. —— "Does Knowledge Require Grounds?" *Philosophical Studies*, vol. 17 (1966), pp. 7–13.

16. —— and CHAMPAWAT, N. "Mr. Clark's Definition of 'Knowledge'." *Analysis*, vol. 25 (1964), pp. 8–9.

17. SOSA, E. "The Analysis of 'Knowledge that *p*'." *Analysis*, vol. 25 (1964), pp. 1–8.

18. WOOZLEY, A. D. *Theory of Knowledge*. London, Hutchinson, 1949.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 4, October 1967

IV. UTTERANCES WHICH INCORPORATE A VALUE STATEMENT

CLIVE BECK

I

ALTHOUGH objectivist theories of value statements often have considerable initial plausibility, they are usually rejected by contemporary philosophers on the ground that, really, they are open to certain fatal objections. Of these objections, the one to which appeal is perhaps most frequently made may be stated as follows: "Every value statement contains one or more elements—emotive, persuasive, imperatival, or the like—which cannot be described as fact-stating, or assessed in terms of truth or falsity. Accordingly, it is impossible for any value statement, taken as a whole, to be an (objective) statement of fact, assessable in terms of truth or falsity." In the present paper, my chief concern is to establish that this objection cannot in fact be sustained, and hence if one wishes to reject objectivism in value theory one must find *other* grounds for doing so.

There are several different varieties of utterance —emotive, persuasive, hortatory, and so on—to which value statements have been assimilated; and the first part of the study will be taken up with describing each of the major ones in turn. Before proceeding with the main discussion, however, it will be necessary to make some remarks about the terms "value statement," and "objective," both of which were used above in the formulation of the objection. To begin with, I must explain why I have chosen to employ the expression "value statement," rather than the more usual expression "value judgment."

The term "value judgment," as it is used in moral philosophy, is somewhat ambiguous. Whereas one might be inclined to think of a value *judgment* as a mental act, or as the content of a mental act, most moral philosophers use the term to refer both to the (mental) value judgment that something is good or right and to the corresponding value statement. This is seen in the fact that the terms "value judgment," and "value statement" are often used interchangeably, even by non-objectivists; and in the fact that value judgments (in this broad sense) are frequently referred to as "speech acts," or described as having been "said," "stated," or "uttered."

In the present paper, the discussion will be carried on not in terms of value judgments (in the broad sense), but in terms of value sattements; and by "a value statement" will be meant "an expressed claim that something is good, bad, right, or wrong (or whatever it may be)." My reason for taking this course is threefold. In the first place, the ambiguity of the technical term "value judgment" would increase, unnecessarily, the complexity of an already difficult field of inquiry. It is true that we wish to solve both the problem of value judgments (in a narrower sense) and the problem of value statements; but there seems to be no reason why we should try to solve both at once. Secondly, since I am here concerned with the relation between value claims and certain non-fact-stating *utterances*, that is, speech acts, it is appropriate that I should talk in terms of value statements. And thirdly, the term "value judgment" has, in recent years, taken on a theory-laden quality such that, in some quarters at least, merely to use the term is to give the impression that one is sympathetic toward non-objectivism. Not wishing to give this impression, I have decided to avoid the term "value judgment."

It might be objected that, in using the term "value statement," I am attempting to gain an unfair advantage. However, to begin with, I cannot see why we should give non-objectivists a monopoly of advantages by always using the expression "value judgment." Again, as I have already mentioned, non-objectivists themselves often use the expression "value statement" as an alternative to "value judgment," apparently without feeling that they have thereby made a concession to their opponents. Further, I have left entirely open the question of the *interpretation* of the term "value statement": if someone wishes to maintain that value statements are not *really* statements at all, even though people popularly describe them as such, or that the class of statements is a very broad one which includes (non-objective) value statements, his theory will not be ruled out by the use of my term "value statement." He is free to give whatever account he thinks fit of "an expressed claim that something is good." Finally, it cannot be denied that in everyday speech people talk of "saying that," "stating that," and "claiming that" something is good or bad, right or wrong. If, then, we are to follow Austin's sound advice, and take ordinary language as our starting point, we must continue to use the terms "value statement" and "value claim" (as well as the term "value judgment," where appropriate) until a theory has been established which shows the illegitimacy of such usage.

The term "objective," as it is used with respect to value statements, is also ambiguous. By describing a value statement as "objective," one may mean either that it is made in an objective manner, with due regard for the relevant facts; or that it is nothing more nor less than an expressed claim *that* something is (objectively) the case. On the one hand, one may be commenting upon the *manner* in which the statement is made; but, on the other hand, one may be talking about the *function* which the statement serves. Now, commands and recommendations can be objective in the first sense: they can be made responsibly, with careful consideration of the relevant facts. Further, the appropriateness of such utterances can be assessed objectively (in the same sense) by outside observers. But it is impossible for commands and recommendations—or emotive, persuasive, and decision utterances—to be objective in the second sense: they simply are not expressed claims *that* something is the case. And, further, it is impossible for an utterance which *contains* an imperatival, hortatory, emotive, persuasive, or decision element to be objective in the second sense; for, taken as a whole, the utterance is something more than an expressed claim, and the description "objective" (in the second sense) does not apply to it. It is clear, then, that what is being denied in the objection under consideration is *not* that value statements may be objective in the first sense—obviously, everyone would agree that they may—but rather that value statements are objective in the second sense. And what, by contrast, objectivists wish to maintain is that value statements are *always* objective in the second sense, since they are always nothing more nor less than an expressed claim that something is (objectively) the case.

Having made these preliminary remarks, I must now describe the various types of non-fact-stating utterance with which value statements have been identified and outline each of the corresponding non-objectivist theories. This will serve not only to clarify the precise nature of the objection under consideration, but also to show that there *is* this common objection to objectivism, inherent in all the main contemporary non-objectivist theories of value. This latter point would be worth establishing for its own sake, for with the increasing sophistication of non-objectivist theories of value there has been a tendency to lose sight of the basic character which they all share, and even to suggest that the difference between objectivist theories and non-objectivist theories is merely a matter of degree.

II

Emotive Utterances

An emotive utterance is a speech act[1] which is an expression of emotion, or an expression of attitude, or both. The order "Shut the door!", for example, said very loudly and in a tone of anger, might be an emotive utterance. And the exclamation "I have sprained my ankle!", made in a certain way, under certain circumstances, may be an emotive utterance. There are several reasons which one may have for voluntarily making an emotive utterance: one may see it as providing some kind of emotional relief; one may use it to arouse sympathy in those who hear the utterance; or one may employ it to persuade the hearers to behave in a certain way. Thus, an emotive utterance is a speech act in its own right, with certain distinctive features and its own rationale.

Some philosophers have maintained that *value* statements are emotive utterances, in the above sense: that they contain an expression of emotion or attitude as an integral element. This view is perhaps best explained by C. L. Stevenson, who states:

> . . . moral terms have a kind of emotional charge, and have it intimately and in their own right. Just as "alas" expresses sorrow and "hurrah" enthusiasm, so "good" and "right" express approval. . . . Mr.

---

[1] I have borrowed J. L. Austin's useful term, "speech act," both because it is sufficiently general for my purposes, and because it is not theory-laden. For Austin's use of the term see, for example, his *How to Do Things with Words* (Oxford, 1962), p. 147.

> Richards calls the force of "hurrah" an "emotive meaning" of the word.... It's one thing to name an emotion and another thing to give direct vent to it; and emotive meaning permits a word to do the latter. Now I am suggesting that "right", "good", and the other evaluative terms have a meaning of this emotive kind.[2]

Further, according to Stevenson, not only do value *terms* have this kind of meaning; value *statements* also have an emotive meaning and, indeed, having an emotive meaning is part of their very nature as value statements. Sevenson says:

> When a man goes from a scientific conclusion to a value judgment it seems to me that he should carefully give notice that the value judgment is *more* than a scientific conclusion.... For the reasons that back up evaluations are . . . less than logically related to the evaluation....[3]

Value statements are "more" than factual statements, according to Stevenson, because they contain an emotive element.

Imperatival Utterances

An imperatival utterance is a speech act which amounts to an order, command, prescription, or instruction,[4] no matter what form of words is actually used. The most typical form of words used, of course, is that of which "Shut the door!" is an instance. But an imperatival utterance may also employ a form of words such as "The door is open," "You ought to have shut the door," "It would be good if you shut the door," or simply "The door." Like emotive utterances, imperatival utterances have a distinctive purpose, namely, to command or instruct someone to do something. Like emotive utterances also, they have distinctive features: they are made in certain types of situations, in a certain conversational context, with a certain tone of voice, with certain facial expressions, with certain emphases, and so on. It is these features which make them distinguishable from other types of utterance in cases where the form of words used is not characteristically imperatival.

Some philosophers have put forward the view that many value statements are imperatival utterances, in the above sense: that they contain an imperative as an integral element. C. L. Stevenson was one of the first to advance this view in recent times; but perhaps R. M. Hare's writings provide the clearest statement of contemporary thinking along these lines. At the beginning of his *The Language of Morals*, Hare asserts that "the language of morals is one sort of prescriptive language";[5] and in its sequel, *Freedom and Reason*, he remarks that, while his views have been the subject of many discussions, he is "obstinate enough to believe that, though they have added much to what (he) thought before, they have not taken much away."[6] Within the class of evaluative statements, Hare makes a distinction between two broad types of statement: statements which commend and statements which prescribe. While evaluative statements which commend and evaluative statements which prescribe fulfill more or less different linguistic functions, depending on the context, the basic purpose which they serve is the same, namely, to teach or affirm, or otherwise draw attention to, not a matter of fact, but how one ought to behave.[7] How one ought to behave may be expressed in the form of standards for choosing between objects, or in the form of rules or principles of action.

Now it might be queried whether this position differs from standard objectivist theories, according to which value statements serve the purpose of indicating, either directly or indirectly, how one ought to behave. There is, however, a considerable difference between the two types of theory; it lies partly in the *manner* in which "how one ought to behave" is indicated, and partly in the kind of *indication* that is involved in each case.

In the first place, whereas on an objectivist view a value statement indicates how one ought to behave by making a *statement of fact*, according to Hare's account a value statement indicates how one ought to behave by issuing a *prescription* to behave in a certain way. To put it crudely, according to the former type of theory, an "ought"-statement refers to a particular action or type of action and *asserts*, "So and so ought to behave like that"; whereas according to the latter type of theory, an "ought"-statement which is genuinely evaluative refers to a particular type of action and *prescribes*, "Behave like that." In the latter case, the *form of words* "One ought to do *y*" or even "*x* is good" may be used; but, nevertheless, the utterance is a prescription to act or to choose in a certain way, not a statement that something is the case.

---

[2] C. L. Stevenson, "The Scientist's Role and the Aims of Education" in I. Scheffler (ed.), *Philosophy and Education* (Boston, 1958), pp. 46–47.

[3] *Ibid.*, p. 50.

[4] In the sense of "instruction" in which one talks of "an instruction" or of "giving instructions."

[5] R. M. Hare, *The Language of Morals* (Oxford, 1952), p. 1.

[6] R. M. Hare, *Freedom and Reason* (Oxford, 1963), p. v. My parenthesis.

[7] *Ibid.*, p. 159. See also R. M. Hare, "Adolescents into Adults" in T. H. B. Hollins (ed.), *Aims in Education* (Manchester, 1964), pp. 59–60.

In the second place, the type of indication is different in each instance. Whereas on an objectivist view a value statement indicates *that* something is the case—for example, *that* someone ought to do something—according to Hare's account the indication is *to* do something, whether to behave in a certain way or to choose a certain object. While on both types of theory, a value statement does in a sense indicate an action or state of affairs, on the former type of theory the value statement indicates something *about* the action or state of affairs, whereas on the latter type of theory it does not.

Persuasive Utterances

A persuasive utterance is a speech act which is an act of persuasion (or at least attempted persuasion). The persuasion may be accomplished (if it is accomplished) simply by making a statement in a certain context; or by using, in addition, special persuasive devices such as a certain tone of voice, certain kinds of words, certain signs of emotion, and so on. The resultant "behavior" may include making certain resolves and having certain emotions as well as performing certain actions. A persuasive utterance is a distinctive kind of action, with a distinctive purpose, namely, to persuade someone to act, think, or feel in a certain way, whether for his own good or for the good of someone else.

Some philosophers have claimed that value statements are persuasive utterances, in the above sense: that they contain an act of persuasion as an integral element. Of these, we may take as an example Stevenson, who makes this comment:

> When I say, "That's wrong!" you tend to feel a little intimidated, don't you? And if, seeing that you have stage-fright, I say, "Go ahead, you'll certainly do a magnificent job!" you're likely to feel more encouraged. A word like "magnificent", in such a context, isn't used to make a prediction but rather to build up an attitude. It gives an emotional push. And there's also an emotional push in "good", "right", and the other evaluative terms.[8]

In *Ethics and Language*, Stevenson says:

> The emotive meaning of a word is the power that the word acquires, on account of its history in emotional situations, to evoke or directly express attitudes, as distinct from describing and designating them.[9]

It should be noted that Stevenson is not simply claiming that words like "good" and "right" can be used in evoking certain emotions and attitudes, that "good" has a conventional significance such that it is a useful word for persuading someone to do something. Rather, he is claiming that an act of persuasion is (typically) *involved in* a value statement: the persuasive act is an integral part of the statement.

Decision Utterances

Decisions and utterances may be linked in various ways. In *deciding* to act in a certain way, or to adopt a certain policy, one may employ an utterance. Usually, of course, people decide *in thought* what they will do and then indicate this decision either in word or in action. But often the *making* of a decision is closely bound up with the uttering of certain words. For example, the uttering of the expression "No, I won't" may be an integral part of the decision not to act in a certain way. Or again, voting "Ay!" may be part of the process whereby one becomes an advocate of a certain proposal. Or again, the making of the statement "I ought to be impartial in my judgment" may be an integral part of the process whereby a judge comes to take up a certain attitude toward a particular case.

Further, in *expressing a decision* which one has already made, one may make an utterance. For example, voting "Ay!" may be, among other things, an expression of a decision one has already made, namely, to vote for a proposal. And the statement "I ought not to be so afraid of responsibility" may be, among other things, an expression of a previous decision to be less afraid of responsibility.

Again, in *informing someone of a decision* already made one may make an utterance. For example, in answer to the question "What have you decided to do?" a person may reply, "I am going to specialize in history," thus indicating what decision he has made. And the statement "It would be wrong for me to give up my music studies" may, among other things and indirectly, serve to indicate a previous (value-neutral) decision to pursue a certain style of life.

---

[8] "The Scientist's Role and the Aims of Education," *op. cit.*, p. 47.

[9] C. L. Stevenson, *Ethics and Language*, (New Haven, 1944), p. 33.

An utterance made in association with a decision in any of these ways I shall call a "decision utterance," the associated decision being understood to be an integral part of the utterance.

It has been suggested by some philosophers of value, notably by Hare and Nowell-Smith, that many value statements are decision utterances, in the above sense: they contain a decision as an integral element. In the course of discussing moral principles, Hare says:

> . . . if pressed to justify a decision completely, we have to give a complete specification of the way of life of which it is a part. This complete specification it is impossible in practice to give. . . . Suppose, however, that we can give it. If the inquirer still goes on asking "But why *should* I live like that?" then there is no further answer to give him, because we have already, *ex hypothesi*, said everything that could be included in this further answer. We can only ask him to make up his own mind which way he ought to live; for in the end everything rests upon such a decision of principle. He has to decide whether to accept that way of life or not; if he accepts it, then we can proceed to justify the decisions that are based upon it; if he does not accept it, then let him accept some other and try to live by it.[10]

Thus, Hare is suggesting that many value statements are expressions of a previous decision—they are a confirmation, extension, and a particular determination of a previous decision to pursue a certain way of life. They are decision utterances of the second kind, mentioned above.

On Nowell-Smith's view, a moral statement of the form "I ought to behave in such and such a way" expresses a decision, indicates what the decision was, and implies that the "I" had good reasons for the decision.[11] He describes "I ought"-sentences as "sentences used for registering decisions,"[12] and says that the utterance "Smith ought to do *y*" when spoken by Smith "expresses a decision."[13] Further, according to Nowell-Smith, the response of "No, I ought not" made to a piece of advice may properly be described as "deciding to reject" the advice.[14] Accordingly, if we do use the words "true" and "false" with respect to "I ought"-statements, we use them in a different sense from that in which they are used in empirical discourse. In empirical discourse, when a man says "That's true" he is endorsing what has been said on the ground (among others) that it corresponds to the facts. But when a person says "Yes, that's true—I ought to do so and so," in a moral context, he is deciding or expressing a decision to behave in the way recommended.[15]

Recommendations and Exhortations

Closely akin to the imperatival employment of value statements is their use in recommendations and exhortations. In recommending a certain textbook, for example, one may simply say "I recommend that you use such and such textbook," implying that one believes the textbook to be a good one and one which should be used. But alternatively one may express this value belief more explicitly in the value statement "This is a good textbook," or "You ought to use this one," leaving it to the timing and tone of the statement to indicate that a recommendation is being made. The resulting utterance is a recommendation which incorporates a value statement.

Again, in exhorting his son to pay more attention to his studies, a man may simply say "*Please* take your studies more seriously," implying that he believes his son's future to be at stake. Alternatively, however, he may express this belief more explicitly in the value statement "You ought to work harder and prepare yourself for your future career!" By virtue of the tone of voice, context, and so on, the utterance remains an exhortation. But it now incorporates a value statement.

Some philosophers of value have maintained that many value statements *are* recommendations or exhortations, in the above sense: they contain a recommendation or exhortation as an integral element. Nowell-Smith, for example, maintains that "you ought"-statements are very often pieces of advice. In such cases,

> it is proper for Brown to say of Jones's advice "that's true" if it would be proper for him to give the same

---

[10] *The Language of Morals, op. cit.*, p. 69.

[11] P. H. Nowell-Smith, *Ethics* (London, 1954), pp. 190 and 261.

[12] *Ibid.*, p. 193.

[13] *Ibid.*, p. 195.

[14] *Ibid.*, p. 195.

[15] *Ibid.*, pp. 196–197.

> advice, which is what he is, in effect, doing; but it is not necessary that he should believe that Jones's advice corresponds to the facts, because pieces of advice are not statements and neither correspond nor fail to correspond to facts.[16]

And Stevenson, to give another example, maintains that "ought"-statements are, among other things, recommendations. Of an argument over whether the school ought to teach Latin he says:

> . . . if we disagree it won't be *just* a matter of your having one set of beliefs about alleged facts and my having another set. We may do that, too, of course; but the main thing is that one of us is *for* teaching Latin and the other is *against* it. We have divergent *attitudes*. . . . and each of us is recommending . . . that his attitude be shared by the other.[17]

III

In the light of the foregoing survey of various types of non-objectivist theory, it should be clear that the conclusion given in the objection under consideration does in fact follow from the premiss: if it is the case that every value statement is an utterance of the kind in question, and hence contains a non-fact-stating element of the kind in question, then it is impossible for any value statement, taken as a whole, to be an (objective) statement of fact, assessable in terms of truth or falsity. For, clearly, expressions of emotion, imperatives, acts of persuasion, decisions, recommendations, and exhortations cannot be fact-stating or true or false. They may be appropriate or inappropriate, successful or unsuccessful, justified or unjustified. But, since they do not state *that* something is the case, they are simply not of the class of phenomena which can (logically) be objective, in the sense in question.

It seems to me, however, that the way is open to an objectivist simply to reject the premiss upon which the objection is based: the proposition that every value statement contains one or more elements of the kind in question. Suppose, for example, that he holds to a rather sophisticated naturalism, statable in some such way as this: The following is a set of logically necessary and sufficient conditions of an object's being valuable for a certain population: (a) the object is satisfying or enjoyable, and/or leads to satisfaction or enjoyment and/or to the avoidance of dissatisfaction or pain, for the population in question; and (b) is practicable, in the long run and on the whole, for that population. Hence, value is empirically determinable; and value statements, that is, statements which either state or imply that something is valuable, are statements of fact, in a broad but legitimate sense of "fact."

The objectivist, then, being in possession of this objectivist theory of value statements, may dismiss the objection in question in the following manner: It is true that, in the course of expressing an emotion, issuing an imperative, performing an act of persuasion, or whatever, one may employ a value statement. But this does *not* mean that, in such an instance, the expression of emotions, imperatives, and persuasive acts is an element *within* the value statement. Rather, the non-fact-stating element is an accidental concomitant of the value statement. The value statement (that is, the expressed value claim) can be distinguished within the total emotive, imperatival, or persuasive utterance in which it is incorporated, and independently assessed in terms of truth or falsity, according to the joint criteria of satisfaction and practicability. While the total utterance obviously cannot be so assessed, this is irrelevant to the question of whether or not the incorporated value statement can be so assessed. Consider, for example, the imperatival utterance, "Your writing is very bad," made in directing a child to improve his handwriting. It seems obvious that the value statement incorporated here can be distinguished from the total utterance in which it is incorporated, and one can ask and answer the question whether or not in fact the writing is very bad (perhaps the teacher has a grudge against the child and is exaggerating the untidiness of the writing). Of course, for the child simply to reply "Yes, that's true" would be inappropriate, for this is not all that the teacher had in mind. But it *would* be in order for the child to say "Yes, that's true; I'll try to improve" and then to set about writing more carefully. For the teacher, in making the utterance, *both* made a statement (hence the "Yes, that's true") *and* issued an instruction to write more neatly (hence the "I'll try to improve," and the corresponding action). Again, consider Stevenson's example of a persuasive utterance: "Go ahead, you'll certainly do a magnificent job!"[18] Stevenson says that the word "magnificent" in such a context is not used to make a prediction but rather to build up an attitude. But surely it is used to do both: the speaker is both saying that he will do a magnificent job (a statement) and attempting to persuade him to think and act in a certain way (an act of persuasion—if successful). The fact that a statement is made is clear if one realizes that an appropriate reply would be "No, you're mistaken; I would bungle it completely." And the fact that two elements are involved—a statement and a total act of persuasion—is brought out if one considers the possible reply: "Yes, I think you're right—here goes!" The first part of the reply is an expression of agreement with the prediction, the second a response to the act of persuasion. Clearly, the statement that he will do a magnificent job can be distinguished from (or within) the total persuasive utterance, and, having been distinguished, can be assessed independently of any assessment of the total utterance. Again, consider the example of a decision utterance of the second type, mentioned above: a man makes the statement, "I ought not to be so afraid of responsibility," as an expression of a previous decision to be less afraid of responsibility. Here, one may point out the value statement so made, and ask whether or not in fact the man *ought* not to be so afraid of responsibility. And in answering this question, the fact that the value statement occurs within a decision utterance is quite irrelevant. What is in question is not whether or not a previous decision was made, of which this value statement is an expression; rather, what is in question is the *truth* of the belief on which the decision was based, namely, that this man (the "I") ought not, for various reasons, to be so afraid of responsibility. And this belief can be assessed without reference to the decision with which it was associated. It would seem, then, that even in cases in which a value statement is incorporated in a non-fact-stating utterance, it is possible to distinguish the value statement within the total utterance. Hence it is a mistake to describe value statements as containing non-statemental elements of the kind in question; indeed, the reverse is the case: emotive utterances, persuasive utterances, and so on (frequently) *incorporate value statements*.

---

[16] *Ethics*, *op. cit.*, p. 197.

[17] "The Scientist's Role and the Aims of Education," *op. cit.*, p. 46.

[18] *Ibid.*, p. 47.

IV

Now, a non-objectivist, having met with such a esponse from an objectivist, might reply in the following manner: It is true that an objectivist *can* take up this stance when confronted by the objection in question, if he *wants* to. But the facts of value discourse clearly suggest that so-called "value claims" are inextricably bound up with non-fact-stating elements. A careful study will reveal that the connection between "value claims" and non-fact-stating utterances in ordinary value discourse is so close that one has good reason to reject the view that objective value statements can be singled out and assessed separately, in terms of truth or falsity.

But *does* one, after a study of the relevant features of moral discourse, have good reason to reject this view? It seems to me that one does not, for at least two reasons. In the first place, one feature of ordinary value discourse which would appear to count against the claim that value statements contain a non-fact-finding element of the kind in question is the sheer number and variety of these elements. We have already noted several such elements, the ones to which attention has been drawn by moral philosophers; and it would seem that many more could be identified, simply by noting the various types of speech act which incorporate an explicit value claim. What is the principle whereby these many and varied speech acts are called value statements? Obviously, it is that they all incorporate an expressed value claim. But, then, what is an expressed value claim? It would be entirely circular to say—as the non-objectivist must—that it is a claim which is associated with one or more of the non-fact-stating elements in question.

Of course, non-objectivists have normally made it a matter of definition that only certain types of non-fact-stating utterance are value statements. For example, Stevenson concentrates upon emotive, persuasive, and "quasi-imperatival" utterances (the term "emotive utterance," for Stevenson, refers to a combination of the three), and rules that statements which do not involve an emotive element can be described as "ethical" only if that term is given "a very broad sense."[19] Again, R. M. Hare makes it a matter of definition that "ought"-statements are imperatival, stating that "I should not say that an 'ought'-sentence *was* being used evaluatively unless imperatives were held to follow from it."[20] But here the non-objectivist is in a trilemma. If he sets up just one non-fact-stating element, as Stevenson does, his theory is open to

[19] *Ethics and Language, op. cit.*, p. 84.

[20] *The Language of Morals, op. cit.*, pp. 164 and 168.

two objections: first, that many value statements simply do not contain *this* non-fact-stating element; and, secondly, that it seems quite arbitrary to choose *this* particular element, from the many one could have chosen. If he mentions two or three non-fact-stating elements, as Hare does, his theory is still open to the two objections just noted (although less obviously) and also to a third: What is the principle of unity among these two or three elements—why, for example, should certain types of commendation, prescription, and decision utterance be lumped together and called by the one name, "value statement"? And if, to avoid the first two objections, the non-objectivist presents a wide range of non-fact-stating elements as the distinctively evaluative components in value statements, his theory is still open, and much more apparently, to the third objection, that is, the one which we began: What is the principle (or the set of principles) of unity whereby these elements are all said to be evaluative?

There is, it seems to me, a second consideration which counts against the claim that the premiss of the objection in question is supported by the facts of value discourse. It may be stated as follows: Often, when people say "x is good," or "y is right," they apparently do so *without* recommending, expressing a favorable feeling, prescribing, deciding, persuading, or anything of the kind. They seem to be simply explaining, reporting, making an observation, describing, or commenting. Take, for example, the teacher who describes a particular student whom he dislikes as a "good artist," without having any favorable feelings toward him; or the historian who says to a physicist that a certain history text is a valuable one (thus explaining why he bought it), without attempting to persuade the physicist to buy the book also, or to take up a certain attitude toward it; or the scientist who mentions to an artist friend that a certain scientific theory ought to be re-examined, without issuing any order or command to the artist. *Something* must be being said in such instances, for people seem to understand such statements, and to consider them worth making. But how can the non-objectivist account for them, while holding to the principle that all value statements contain a non-fact-stating element of the kind in question? He cannot allow that a genuine value statement is made in such instances, for apart from the fact that this would involve an admission of the incompleteness of his theory, it would also raise the obvious query: If a value statement may be made without an associated expression of emotion, act of persuasion, issuing of an imperative, or whatever, does this not suggest that such non-fact-stating elements are always only an accidental concomitant of a value statement, and not an integral component within the statement? But, on the other hand, it is difficult to see how the non-objectivist can explain away such cases.

Stevenson, in an attempt to overcome this difficulty, introduces the notion of an "inactive emotive meaning," in terms of which he explains how it is that a value statement can have an emotive meaning and yet not express an emotion or attitude. But this will not do, for while it is true that we talk of a *word* or *sentence* as meaning p and q and r (and even then, to be more correct, we should say that the word or sentence *can* mean either p, or q, or r), regardless of how it is used in any particular situation, yet a *statement* has no other meaning than that which it has in the particular context (or kind of context) which makes it a statement. Hence, while the *word* "good" may be said, without qualification, to have an emotive meaning, even if it is not *always* used to express an emotion or attitude, a *statement* in which the word "good" is employed only has an emotive meaning if, in the context (or type of context) in question, it is used to express an emotion or attitude. Or, to put the point more strongly, if the statement in which the word "good" is employed is not being used to express an emotion or attitude, then the word "good" as it is used in *that* statement does not have an emotive meaning. To say that it has an "inactive emotive meaning" can amount to no more than saying that in some contexts it has an emotive meaning; and this obviously gives no support to Stevenson's claim that whenever the word "good" is used, there is an emotive meaning present. Stevenson, then, does not succeed in meeting his second objection to the claim that all value statements contain a non-fact-stating element of the kind in question; and it seems to me that neither this objection, nor the first one which I mentioned, can in fact be met.

V

The conclusions which I wish to draw on the basis of the foregoing discussion are quite limited in scope. I cannot claim to have provided a refutation of the non-objectivist theories we have been considering, as I have presented only a few pieces of evidence against them; and I certainly cannot claim to have established an objectivist theory of value statements. However, in the first place, it is undeniable that the considerations raised in the preceding section do count against non-objectivist theories of the kind in question. And, secondly, it is clear that the particular objection with which I have been concerned in this paper is itself beset by so many problems that it does not present a difficulty for objectivist theories of value statements. If the premiss on which the objection is based could be shown to follow from some well-established theory, the objection would have some force. But since no such theory is at hand, and since the facts of ordinary value discourse provide no support for the premiss, and indeed suggest that it is false, one may dismiss the objection and conclude that if objectivism in value theory is to be rejected, *other* grounds for the rejection must be found.

University of Toronto

Received October 12, 1966

American Philosophical Quarterly
Volume 4, Number 4, October 1967

V. THE EQUIVALENCE CONDITION

HOWARD SMOKLER

WHAT are the characteristics of a correct concept of confirmation? Inductive logicians have made use of a variety of techniques in attempting to answer this question. The statement of conditions of adequacy, and the use of these conditions to determine whether or not a putative definition is acceptable, is one of the principal instruments in their toolbox.[1] In general, these conditions do not uniquely select a relation which can be shown to fulfill them; rather they select a class of such relations. Since our intuitions regarding the relation of confirmation are far from sharp, investigation may force us to modify or even abandon a condition which seemed at first the quintessence of reasonableness.

Most of the conditions of adequacy which have been stated link the confirmational relation to a semantic relation which holds between statements. Thus, an early form of the equivalence condition was formulated in the following way:

> (8.22) Equivalence Condition: If an observation report confirms a hypothesis H, then it also confirms every hypothesis which is logically equivalent to H.[2]

This links the inductive relation of "confirmation" with the semantic relation of "logical equivalence." The latter notion, in turn, is defined relative to a specific logical language and system. The condition reflects a fundamental assumption which most students of the subject make. This assumption is that a system of confirmation-theory presupposes an underlying deductive logic (associated with a specific language) for which standard metalogical notions are defined.[3] The character of the linkage is specified, at least partially, by indicating the relation which certain semantic notions bear to the confirmation relation, a relation which is itself generally taken to be of a metalinguistic character.

What are to be taken as conditions of adequacy is a delicate question. Conditions suggested by reflections on semantic questions; conditions which permit a simple and advantageous statement of the confirmation relation; conditions suggested by reflections upon the way we make use of the concept of confirmation in ordinary life and in our scientific activities; these are some of the sources. These conditions are all taken to be true statements about the relation of confirmation to some semantic notion. But the formal relationships between the conditions and the consequences to which they lead also serve as a source for their modification.[4] With such heterogeneous sources, it is only natural to suppose that some of the conditions are more easily modified or given up than are others. And this is precisely the situation. Of all the conditions which have been suggested, the one most tenaciously upheld has been a generalized form of the equivalence condition. It is not difficult to understand why this has been the case. It accords very well with positions held most emphatically in the theory of meaning and epistemology; in addition, it is in accord with the formal apparatus of currently held explicata of the notion of confirmation.[5]

I want to argue against the acceptance of this generalized equivalence condition, at least in the form in which it is presently stated. To do this, I first examine the plausibility of some of the arguments which have been offered for its acceptance. Next, I state an underlying assumption which lends such a strong plausibility to this condition, in the minds of its upholders. I argue for a modification of this assumption, and offer, in conclusion, a sketch of an alternative point of view towards the condition which I believe should be adopted.

---

[1] Carl G. Hempel was the first to introduce the use of conditions of adequacy into the study of confirmation theory. He did this in his two classic papers, "A Syntactical Definition of Confirmation," *The Journal of Symbolic Logic*, vol. 8 (1943), pp. 121–143, and "Studies in the Logic of Confirmation," *Mind*, vol. 54 (1945), pp. 1–26, 97–121. The latter article is reprinted in Hempel's recent book, *Aspects of Scientific Explanation* (New York, 1965).

[2] Hempel, "Studies in the Logic of Confirmation" in *Aspects of Scientific Explanation*, *op. cit.*, p. 31.

[3] This assumption is clearly to be found in Hempel's work, especially in the first article cited in footnote 1. For another author who shares this point of view, see Wesley C. Salmon, "On Vindicating Induction," *Philosophy of Science*, vol. 30 (1963), pp. 252–261, especially p. 258.

[4] The need for constant modification of these conditions is underlined by the investigations of Rudolf Carnap, *Logical Foundations of Probability*, (2nd ed., Chicago, 1962), pp. 468–478; Nicholas Rescher, "A Theory of Evidence," *Philosophy of Science*, vol. 25 (1958), pp. 83–94; Wesley C. Salmon, "Consistency, Transitivity, and Inductive Support," *Ratio*, vol. 8 (1965), pp. 164–169, as well as of a number of other writers.

[5] For example, some version of the equivalence condition is presupposed in accounts of metrical confirmation which satisfy the axioms of conditional probability.

I. The Generalized Equivalence Condition

The generalized equivalence condition, which is an outgrowth of Hempel's equivalence and inverse equivalence conditions, may be roughly stated in the following way:[6]

> (1a) Given a language L and a standard (quantificational) logic of that language, for all confirmation-contexts of L, logically and definitionally equivalent statements are substitutable for one another, *salva veritate.*
>
> (1b) Given two languages, L_i, L_j which fulfill certain conditions,[7] there exists a homeomorphic translation function over the relation of confirmation which carries every statement in language L_i of the form
>
> S_i confirms S_j
>
> into a statement of the language L_j, of the form
>
> S'_i confirms S'_j
>
> *salva veritate.* (S'_i is a translation of S_i, as is S'_j of S_j. The translation function treats any sentence S_k which is logically equivalent to a translation of S_i as a translation of S_i; this is the homeomorphism.)

The equivalence condition in both its parts is a principle of the identity of indiscernibles. The formal language or languages for which the confirmation function is defined is generally taken to be an applied quantificational language; the corresponding logic is quantification theory.[8] Confirmation relations are generally taken to be stated in the metalanguage of the language for which the relations of logical and definitional equivalence hold.[9] Yet there is no uniformity of opinion on this question. When explicit thought has been given to the problem, a metalinguistic formulation has been preferred. The reason for this is that for a language in which the arguments of confirmation relations were sentences, all confirmational-contexts would be non-extensional. And furthermore, as Carnap has suggested, the language would contain logical modalities. Given the preference for extensional languages and the desire to restrict semantics to the theory of reference, it is not difficult to see why this course has been taken. But the question of formulation is still an open one, and later in this paper its relation to the equivalence condition will be discussed.

Clearly, the condition as it stands is not an *a priori* truth. Its plausibility rests upon certain commonly held assumptions in semantic theory. Investigation of these assumptions requires that they be drastically modified if the condition is not to be abandoned.

II. Defense of the Condition

The best way to uncover the assumptions to which we have just referred is to investigate some of the arguments for the plausibility of this condition which have been offered. The first argument, in a restricted form, is that of Hempel. It has two parts which are summarized in the following quotation:

---

[6] Hempel stated a more narrow set of conditions (which amount to a special case of this one) for a specified language L in the two papers which are cited in footnote 1. Wesley C. Salmon and John R. Wallace have independently stated the generalized equivalence condition. Salmon first stated the condition in his paper, "Vindication of Induction" in *Current Issues in the Philosophy of Science*, ed. H. Feigl and G. Maxwell (New York, 1961), pp. 245–264; "Inductive Inference," *Philosophy of Science: The Delaware Seminar* II, ed. B. Baumrin (New York, 1963); "Use, Mention, and Linguistic Invariance," *Philosophical Studies*, vol 17 (1966), pp. 13–18. John R. Wallace has stated the condition in its generalized form in "Goodman, Logic, Induction," *The Journal of Philosophy*, vol. 63 (1966), pp. 310–328. Salmon calls the condition the criterion of linguistic invariance. Wallace has pointed out its connection with Hempel's conditions.

[7] Wallace discusses some of these conditions in his article, especially on pp. 311–313.

[8] This is explicitly assumed by Carnap and Hempel. Wallace assumes that quantificational logic is the logic with which confirmation must be linked; see p. 311 of his article. His investigations do raise doubts in his mind as to whether this is the proper underlying logic and language for confirmation theory; see pp. 327–328 of his article.

[9] Carnap in the *Logical Foundations of Probability*, *op. cit.*, pp. 279–284, discusses two alternative formulations of the confirmation relation. In the first, names of sentences are taken as arguments; in the second, sentences (which express propositions) are the arguments. The former formulation is metalinguistic. Carnap favors it because he claims that it permits the language in which the relation is expressed to be extensional and the underlying logic to be non-modal. Recently, Carnap has decided to take as arguments for confirmation relation propositions interpreted as sets of models. What effect this will have upon the underlying logic, I do not know. Hempel's formulation of the relation is metalinguistic.

> (1) Fulfillment of this condition makes the confirmation of a hypothesis independent of the way in which it is formulated; and no doubt it will be conceded that this is a necessary condition for the adequacy of any proposed criterion of confirmation. Otherwise, the question as to whether certain data confirm a given hypothesis would have to be answered by saying: "That depends on which of the different equivalent formulations of the hypothesis is considered"—which appears absurd.
> (2) Furthermore—and this is a more important point than an appeal to a feeling of absurdity—an adequate definition of confirmation will have to do justice to the way in which empirical hypotheses function in theoretical scientific contexts such as explanations and predictions; but when hypotheses are used for purposes of explanation or predictions, they serve as premises in a deductive argument whose conclusion is a description of the event to be explained or predicted. The deduction is governed by the principles of formal logic, and according to the latter, a deduction which is valid will remain so if some or all of the premises are replaced by different but equivalent statements; and indeed a scientist will feel free, in any theoretical reasoning involving certain hypotheses, to use the latter in whichever of their equivalent formulations are most convenient for the development of his conclusions.[10]

Hempel clearly assumes that if two sentences S_i and S_j which are formulated in a language L of the appropriate type are logically equivalent, then they express the same fact. Salmon seems to express the same presupposition when he declares:

> If our logical systems are to reflect the fact that evidential relations do not depend upon linguistic conventions, then the rules of inference admitted into these systems must be invariant with respect to purely linguistic transformations of premises and conclusions.[11]

Since "purely linguistic transformation" refers to logical equivalence of sentences, we may assume that Salmon, too, is thinking of the logical equivalence of sentences as a necessary and sufficient condition of their synonymy. Furthermore, there is another presupposition which both authors share, an assumption which is enthroned in the condition. It is this: logical equivalence is a mark of synonymy in confirmation contexts.

Hempel and Salmon have argued in the following way. The condition of equivalence is an obvious one, since it is clear that logically equivalent sentences have the same meaning though different forms. In addition, Hempel argues, confirmed sentences are a necessary condition for the explanation of singular sentences. But in explanatory contexts all logically equivalent sentences are treated as indiscernibles; therefore all the equivalents of confirmed sentences must be confirmed. Salmon adds an argument of his own. The relation of confirmation is one that holds between non-linguistic entities.[12] This fact implies the general equivalence condition, for the denial of the condition would make the truth of statements of confirmation dependent upon the character of the language for which the confirmation relation was stated, and this would be an intolerable situation.[13]

None of these arguments is conclusive. Their plausibility is based upon the assumption that some criterion of synonymy should operate so as to permit us to treat certain sentences as indiscernible from one another; careful examination of the situation reveals that if this is the case, the criterion of logical equivalence will not do the job. Thus, Hempel's first argument is based upon an appeal to an intuition of absurdity regarding any contrary of the equivalence condition. But in the form in which it is presently stated, it is plausible to accept a contrary of the equivalence condition.

Both the paradox of the ravens and Goodman's conundrum regarding grue and bleen are arguments against the equivalence condition in its present form. In its original form, we are presented with the following two true statements which form the basis of the paradox of the ravens:

---

[10] Carl G. Hempel, "Studies in the Logic of Confirmation" in *Aspects of Scientific Explanation*, *op. cit.*, p. 13.

[11] Wesley C. Salmon, "Use, Mention, and Linguistic Invariance," *Philosophical Studies*, vol. 17 (1966), p. 13.

[12] Thus Salmon writes, "The validity of a deductive argument does not come into being with the statement of that argument ... The same considerations, it seems to me, apply to inductive correctness. Even if the universe contained no thinking or language-making creatures, certain facts would constitute inductive evidence for other facts" (*ibid.*, p. 13). Incidentally, Salmon's position lends weight to the hypothesis that he considers confirmation as an object-language notion. Whether he would be prepared to accept some of the consequences of this decision is another question.

[13] Recently Salmon has offered some other arguments for the generalized equivalence condition. He argues that the condition really is a consistency requirement and secondly that its violation would lead to irrational behavior, i.e., to the making of a set of bets which are incoherent. I do not find Salmon's arguments compelling. They rest upon the assumption that a statement of confirmation in L_1 is synonymous with a statement of confirmation in an alternate language L_2 if the only way that the statement in L_1 differs from that in L_2 is that it contains as sub-sentences, sentences which are the logical equivalents of the appropriate sub-sentences of the sentence in L_2. This is the very point at issue, and the paradoxes adduced in this paper are arguments against this identification. The argument is found on pp. 236–237 of his recent article "The Foundations of Scientific Inference" in *Mind and Cosmos*, ed. by R. Colodny (Pittsburgh, 1966; University of Pittsburgh Series in the Philosophy of Science, vol. III).

(1) a_i is a raven and a_i is black CONFIRMS all ravens are black.

(2) a_i is a raven and a_i is black DOES NOT CONFIRM all non-black objects are non-ravens.[14]

These two statements are incompatible with the present formulation of the equivalence condition; this can as well be taken as an argument against the condition as against the truth of the statements. But similar considerations can be seen when we consider the paradox of grue and bleen. In that case the evidence class can be stated in two definitionally equivalent ways:

$$E = \begin{cases} a_1 \text{ is an emerald \& } a_1 \text{ is green \& } a_1 \text{ is examined before } T \\ \dots\dots\dots\dots\dots\dots \\ a_n \text{ is an emerald \& } a_n \text{ is green \& } a_n \text{ is examined before } T \end{cases}$$

$$E' = \begin{cases} a_1 \text{ is an emerald \& } a_1 \text{ is grue \& } a_1 \text{ is examined before } T \\ \dots\dots\dots\dots\dots\dots \\ a_n \text{ is an emerald \& } a_n \text{ is grue \& } a_n \text{ is examined before } T. \end{cases}$$

Yet clearly

(1′) E' CONFIRMS all emeralds are grue.

(2′) E DOES NOT CONFIRM all emeralds are grue.[15]

This argument shows that the equivalence condition in its present form is not at all evident, as some authors are apt to claim.

Hempel offers a second argument as well. It can be summarized in the following way. Hypotheses which are confirmed by the evidence enter into the explanation or prediction of singular statements. Deductive-nomological explanations are defined in terms of a certain set of consequence relations holding between sentences. Hempel claims that the equivalence condition must hold for sentences which stand in the explanatory relation with one another. If this is the case, then if a confirmed hypothesis H explains some singular sentence S. any logical equivalent of H explains S. If the equivalence condition does not hold for sentences in confirmational contexts, then H may be confirmed while its logical equivalent is not confirmed. This situation, Hempel claims, is insupportable. It therefore provides an argument for the acceptance of the condition. Reflection upon this argument reveals that it is not at all persuasive for two reasons. First of all, it is not clear that the generalized equivalence condition is fulfilled in explanatory contexts. As Hempel himself has stressed, all hypotheses must be lawlike statements in order to be part of an explanation. But the logical equivalent of a lawlike statement need not be a lawlike statement, at least if a purely semantic definition of lawlikeness is given.[16] Therefore, the fulfillment of the equivalence condition in explanatory contexts is problematic. And even if this condition were fulfilled, there is no reason to suppose that scientists could not make (and would not make) pragmatic distinctions between logically equivalent hypotheses, employing a confirmed hypothesis in explanation, but rejecting its non-confirmed logical equivalent.

An argument based upon somewhat different grounds is offered by Wallace. He criticizes Goodman's system of confirmation theory by showing that if one assumes both this condition and standard quantificational theory as the underlying logic of the language to which confirmational theory is applicable, then Goodman's theory faces insuperable difficulties. Wallace never argues directly for the acceptance of the condition, and this leads one to believe that he thinks of it as an *a priori* condition whose contraries would be absurd. But he does provide a framework for understanding the relation between confirmation theory and deductive logic, in the section of his article entitled "Logical and Definitional Equivalence" (*op. cit.*, pp. 324–327) and this may be taken as an explication of the character of this *a priori* condition. Wallace's line of thought can be sketched in this way: consider two languages (either constructed or natural) and the conditions for communication between persons using the two languages. Assume furthermore that the languages are radically different from one another. In order to communicate, two people who use the two languages must be able to translate from one language into another. Such a process is intimately related to the intuitive notions of synonymy, as Quine shows in his chapter on Translation and Meaning in *Word and Object.* Wallace, following Quine, wants to see what conditions of invariance of translation must be fulfilled in order that communication be possible. Quine had recognized some types of contexts in which there is an invariance; thus, for example, observation-sentences in different languages are stimulus-synonymous.[17] Wallace wishes to extend this invariance to pairs of sentences which stand in the confirmation relation to one another. The condition of equivalence stated earlier in this paper is a statement of this invariance. He sketches an account of the conditions under which mutual comprehension of two radically different languages is possible by the user of each of the languages; alternatively, mutual translation from one language to another is possible. Recursive devices for specifying meaning, for example, must be available in both languages (*op. cit.* pp. 325–326). Not all contexts, however, permit the treatment of logically and definitionally equivalent statements or theories as indiscernibles when translations are effected. Thus:

---

[14] The confirmation relation has been taken here as a term in the object language, but it could clearly be given a metalinguistic formulation.

[15] This has led some authors to reject the equivalence condition explicitly. Goodman does this in his reply to Wallace, "Comments," *The Journal of Philosophy*, vol 63 (1966), pp. 328–331, especially p. 331. Scheffler has shown that Goodman's notion of confirmation does not satisfy the equivalence condition in his book, *The Anatomy of Inquiry* (New York, 1963), pp. 287–289, and I have shown that it does not satisfy the inverse equivalence condition in my review of Hempel's *Aspects of Scientific Explanation* in *Synthese*, vol. 16 (1966), pp. 110–122. Rescher has also rejected the equivalence condition for confirmation-theory. He claims that the contrapositive of a lawlike generalization is not substitutable for the generalization in inductive contexts, and he makes use of this observation to deal with Hempel's paradox. See his recent book, *Hypothetical Reasoning* (Amsterdam, 1964), especially Appendix 3.

[16] At the end of a discussion of a semantical criterion of lawlikeness Hempel offers the following definition: A sentence S is lawlike if and only if it does not have a finite number of instances. He then shows that upon this criterion, the logical equivalent of a statement which is lawlike is not itself lawlike. For consider the sentence

(1) All F are G.

and its logical equivalent

(2) Anything that is F but not G is both G and not G.

Clearly the second sentence has no instances. This may lead one to a pragmatic criterion of lawlikeness of the kind that Goodman suggests in *Fact, Fiction, and Forecast* (second ed., Indianapolis, 1965). Projectible hypotheses would be lawlike, but for the reasons which we have already discussed, the fact that H is projectible in no way guarantees that a logical equivalent of H is projectible. Hempel's discussion of lawlikeness is contained in *Aspects of Scientific Explanation*, *op. cit.*, especially p. 341.

> In other words, which notions satisfy the principle of indiscernibility of definitional equivalents? Logical consequence, as we expect, corresponds exactly. Effects caused by poems memorized in childhood, by a disposition to pun, or by a weakness for words of Anglo-Saxon origin, as we expect, are represented as being idiosyncratic. *And so are confirmation, lawlikeness, contrariety, counterfactual connection, instancehood (in the sense of chapter iv of FFF), support, violation, evidence class, exhaustion (as said of hypotheses), and projectibility.*[18]

On the face of it, this is a damaging admission for Wallace to make. The equivalence condition fails for confirmational contexts; it was Wallace's original claim that the condition must hold for confirmational contexts. Perhaps he is claiming that these contexts, when defined in the way suggested by Goodman, do not satisfy this invariance condition. An alternate account of confirmation or of the underlying logic might permit the fulfillment of this indispensable condition of communication. If this is Wallace's claim, he has not demonstrated it. *Prima facie*, communication (and translation) of one language into another is possible even if all contexts are not such as to permit of invariance in translation. This may entail, as a consequence, that these contexts are language dependent; a consequence that Salmon would wish to deny. Though I can see this as an intuitively disagreeable consequence, I do not believe that it would make inductive logic a travesty. In fact, Goodman's system of inductive logic is precisely of this character.

There is another consideration which Wallace might raise. He could claim that a theory of meaning involves a recursive account of the meaning of any expression of a language, and that the failure of the equivalence condition in confirmational contexts would make it impossible to give such a recursive account of the meaning of such contexts.[19] This is an interesting idea, but it must be developed in more detail before its truth or falsity is determined.

---

[17] W. V. Quine, *Word and Object* (New York, 1960), ch. II, especially sect. 15.

[18] Wallace, *op. cit.*, pp. 326–327, italics mine.

[19] That Wallace had this position at the back of his mind is suggested by his discussion on p. 325 of his article, *op. cit.* Given the present state of the theory of meaning, the acceptance of such a recursive condition by Wallace should lead him to accept confirmation as an object-language notion. In his article, "Theory of Meaning and Learnable Languages," *Proceedings of the 1964 International Congress for Logic, Methodology, and Philosophy of Science*, ed. Y. Bar-Hillel (Amsterdam, 1965), especially pp. 388–390 Davidson has shown that there are great difficulties in reconciling a recursive account of meaning with the employment of metalinguistic contexts in which the names of object-language expressions occur. This would necessitate a change of structure in the underlying logic. For quantificational logic is extensional in character. Adding statement operators like the confirmation

Arguments which have been given for the equivalence condition in its present form are not at all conclusive. Rather the condition in its present form is clearly false. This can lead one to abandon the condition completely (as Goodman does) or to modify it (or the underlying logic with which it is associated) so as to make it true. In either case, the problems which must be solved are those in the theory of meaning.

III. The Theory of Meaning

The condition of the indiscernibility of logical and definitional equivalents is a characterization of synonymy. This explains why the authors who have been mentioned offer such a paucity of arguments, and why they treat the condition as self-evident. In particular, this assumption is compatible with and underlies Hempel's view that, otherwise, confirmation would depend upon the formulation of the hypothesis and not upon its content. The same thing can be said for Salmon's argument. And the same assumption underlies Wallace's argument, though he makes use of the device of radical translation to underscore the more sophisticated treatment that synonymy requires when we consider Quine's writings upon the subject.

Viewed in this light, several paths are open to us. The first is to give up the attempt to state a condition of synonymy within confirmation theory. Goodman has taken this path; it is perfectly consistent with his general position regarding synonymy.[20] This position seems like strong medicine; until it is conclusively shown that the notion of synonymy is of no philosophical value, it should not be abandoned. Thus, there must be a search for an alternative formulation of the condition of equivalence (or for some other strategy) which will provide a proper criterion of synonymy with confirmation theory. Several strategies are available.

These strategies can be considered as attempts to surmount the difficulties occasioned by the following fact: in confirmational contexts, if the logic assumed is quantificational logic, then logical equivalence will not be a mark of a sentence synonymy. For a necessary and sufficient condition of sentence (and theory) synonymy is that substitution of synonymous sentences (or theories) within any context takes place *salva veritate*. The paradox of the raven as well as that of "grue and bleen" show that when logical equivalence is taken as a mark of synonymy, the criterion for synonymy fails. Either the underlying logic of the language (and correspondingly the language) must be changed, or some "tighter" criterion of synonymy than logical and definitional equivalence must be devised, if the equivalence condition is to be preserved. There is still another possibility: to proceed in a way similar to the way in which epistemic notions like "knowledge" and "belief" have been treated as modal operators. A successfully devised system of modal logic would incorporate in its rules of transformation some version of the synonymy condition.

In fact, it is not surprising that the present notion of logical and definitional equivalence will not do as an explication of synonymy. For such an explication of synonymy is intimately connected with the verificationist theory of meaning. Quine had indicated the connection, though in a somewhat different context.[21] The verificationist theory of meaning considers the meaning of a sentence to be its manner of confirmation. Two sentences are synonymous if they have the same manner of confirmation. They have the same range of verifying experiences. Therefore, logically equivalent sentences have the same method of confirmation. Thus it follows that they are synonymous. Quine has argued, convincingly, I believe, that the verificationist theory of meaning is a dogma; it follows that all the concepts based upon it have the flavor of the dogmatic. However insufficient as a theory of meaning, it still provides a plausible basis for the criterion of synonymy which we have been discussing. And when that criterion has independently been shown to be wanting, this

---

relation would change all contexts in which they occur into non-extensional contexts, and this would require a corresponding change of the underlying logic of the language. This possibility is considered later on in this essay. Incidentally, Quine has shown that for any language in which logically equivalent sentences are interchangeable *salva veritate* (the equivalence condition and one in which the logic of classes is at hand), any form of statement composition, other than truth-functional, is referentially opaque. Clearly, confirmational contexts are not truth-functional. Therefore, in such a language, they would be referentially opaque. See W. V. Quine, *From a Logical Point of View* (Cambridge, 1961), p. 159.

[20] N. Goodman, "On Likeness of Meaning," *Analysis*, vol. 10 (1949–50), and "On Some Differences about Meaning," *Analysis*, vol. 13 (1952–53). Goodman even raises doubts about the correctness of the equivalence condition in contexts involving deductive logic only. See his reply to Wallace cited in n. 15.

[21] W. V. Quine, *From a Logical Point of View*, *op. cit.*, ch. II.

casts doubt on the verificationist theory of meaning.[22]

The situation which we have described may be summarized in the following way. Inductive logic is the theory of a certain relationship (or relationships) which exists between pairs of sentences of a language. In general the languages for which confirmational relations have been defined are extentional languages, i.e. languages in which, for all contexts the substitution of extensionally equivalent expressions does not change the truth-value of the context. Some confirmational contexts are not extensional, when the extensions of sentences are taken as truth-values, confirmational relations are generally taken to be stated in the metalanguage. (This requirement conflicts with other demands which would make it more plausible to state the relation in the object-language; these conditions have been discussed in the previous section.) Within these extensional languages, extensional equivalents cannot be taken as synonymous. Considerations such as those discussed in the preceding paragraphs have led writers to take logical equivalence as the proper mark of synonymy. But this conflicts with a reasonable requirement on synonymy: substitution of synonymous expressions must not change the truth-value of the context in which the expression occurs.[23] The paradoxes of confirmation show that logical equivalence is not a mark of synonymy. Under these conditions, a way of accommodating the following assumptions must be found:

(1) In language L whose underlying logic is quantificational, sentence-synonymy can be explicated by the logical equivalence of sentences. This notion can be extended to theories.

(2) In confirmational contexts (whose arguments are the names of sentences in L) the substitution of logically equivalent sentences for one another (more precisely of their names) may change the truth-value of the context.

(3) Confirmational contexts should preserve their truth-value under substitution of synonymous sentences for one another.

Several strategies for reconciling this set of incompatible assumptions have been mentioned above. In all cases, the problem to be solved is fundamentally the same. Upon the usual definition of synonymy as logical equivalence, the confirmational relation is not invariant under substitution of synonyms. A characterization of synonymy (or indiscernibility) must be given for which confirmation is invariant under substitution of synonyms. Such a solution might be given at either the object or the meta-linguistic level.

One way to do this would be to change the underlying language (and its associated logic) in such a way that these crucial statements which are logically equivalent in quantificational language are no longer logically equivalent. Synonymy would still be defined in terms of logical equivalence. Since another underlying logic was being used, synonymy would have a different meaning. I suspect that the language employed would turn out to be a modal one. Confirmation would still be a meta-linguistic notion defined for sentences in the language, and the equivalence condition would still hold in its present form. Thus assumption (1) would be modified.

Another possible strategy of approach would be to make the relation of synonymy a tighter one than that of logical equivalence. The underlying logic and language would be kept fixed, but more stringent conditions than logical equivalence would be taken as a mark of synonymy. Carnap, for example, has proposed taking intensional isomorphism as an explicatum for synonymy. Whether or not this is a successful maneuver, it suggests a line of approach. Whatever condition is accepted, it would have to guarantee that the statement in terms of which the various paradoxes of confirmation are formulated are not synonymous with one another.

---

[22] The fact that the paradox of the ravens can be considered as an argument against the verificationist theory of meaning was pointed out to me by Dr. Dagfinn Føllesdal.

[23] The referential opacity of a context has frequently been taken as a mark of its non-extensionality. Since practically all metalinguistic contexts contain the names of sentences (formed by quotation marks applied to the sentences), most are referentially opaque upon a plausible definition of referential opacity. For, as Quine points out in "Three Grades of Modal Involvement," *Proceedings of the 11th International Congress of Philosophy*, vol. 14 (Amsterdam, 1953), pp. 65–81. "Quotation is the referentially opaque context *par excellence*" (p. 67). In particular, this is the case for confirmational contexts when they are interpreted metalinguistically.

The real logical structure of confirmational contexts is not revealed by phrasing them in metalinguistic idiom. All confirmational contexts are referentially opaque merely because they contain a name of a sentence, i.e., they contain a quotation-context. But confirmational contexts are referentially (as well as extensionally) opaque on independent grounds, and these can be seen most clearly when an object-language formulation is given to them. In any case, the doctrine of quotation-contexts both functioning as names and being referentially opaque is not coherent.

Still a third position could be taken. Confirmational contexts are obviously non-extensional and referentially opaque. Confirmation should be treated directly as a modal operator upon pairs of sentences. As in all modal logics, we must deal with the problem of what kinds of substitutions are permissible within the language. We must directly face the problem with which we have been struggling. Furthermore, if such a modal formulation is given, the incoherencies associated with a metalinguistic formulation of the position are avoided, and confirmation is treated as the operator on sentences which we commonly consider it to be. In addition, the modal character of the notion of confirmation is clearly shown, and this is a fact which is obscured by most present formulations. Though there are very great difficulties associated with the carrying out of such a program, it seems to me the most interesting and philosophical of the various strategies which I have outlined.[24]

To summarize the argument of this paper: the plausible requirement of treating sentences which are logically equivalent in a standard system of logic as indiscernibles has been advanced on the basis of a certain theory of meaning. Consideration drawn from the investigation of inductive logic (the paradoxes of confirmation) demonstrate that such a condition is not acceptable as it is usually stated. We are forced to reconsider assumptions of a semantic character which are involved in the formulation of the relationship between confirmation theory and the underlying logic of the language for which the confirmational relation is defined. Various ways of overcoming some of the apparent incompatibilities between acceptable assumptions are considered, but no final conclusion is reached regarding the proper solution. However, the modal character of the confirmational relationship is underscored by these investigations.

C. W. Post College of Long Island University

Received March 8, 1967

---

[24] The greatest difficulty associated with a modal formulation is in providing a satisfactory semantics for the system. Great steps toward a solution of this problem are to be found in the works of Jaakko Hintikka, Dagfinn Føllesdal, and Stig Kanger. See Jaakko Hintikka, "The Modes of Modality," *Acta Philosophica Fennica*, fasc. 16 (1963), pp. 65–82; Dagfinn Føllesdal, *Referential Opacity and Modal Logic* (Unpublished Doctoral Dissertation; Harvard University, 1961); and Stig Kanger, *Provability in Logic* (Stockholm, 1957). That probability is a modal context is an old idea. In "A New System of Modal Logic," *Logical Studies* (London, 1957), G. H. Von Wright has set forth a new modal system which is dyadic and whose metricized form is a calculus of probability. It probably could be employed to formulate qualitative confirmation as well. At present, only sentential variables are employed; the system would have to be extended to allow quantification.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 4, October 1967

VI. "MIRACLE" AND "PARADOX"

ALASTAIR McKINNON

I

THE Christian religion at least has been tied up with belief in the reality of certain events which are very far from our ordinary experiences and conceptions. These events the Church has traditionally conceived as miracles and paradoxes. This paper deals with these terms regarded as descriptions of events. It is, in fact, an attack upon them. More precisely, it is an attempt to show that the traditional or supernatural senses of these terms cannot consistently name or describe any real or alleged event.

It is true that these terms have a sense which is both properly religious and logically defensible. It is also true that this sense has sometimes been to the fore. Nevertheless both terms are now generally understood in a supernaturalist sense or, rather, in a variety of such senses. It is against these alone that the present comment is directed.

Three observations may be made at the outset. The first is that each of these terms has both an expressive and a descriptive use. The line between these uses is not easy to draw in theory and only rarely possible to maintain in practice. But there is no real difficulty in grasping the distinction. For example, in referring to an event as a miracle I might mean to call attention to the fact that it did not square with my previous conception of nature. Or I might intend to assert that it was discrepant with the real or established order of nature. The former would be the expressive and the latter the descriptive use of this term. It is with the latter use alone that this paper is concerned.

The second observation concerns a particular aspect of the traditional or supernatural definitions of these terms. Historically these have been qualified, at least implicitly, by some phrase such as ". . . and owing to some divine agency." Despite this, I propose to discuss these senses without reference to this qualification. Later I hope to justify this procedure but for the moment content myself with the observation that it is irrelevant to the real point at issue. Or, better, that it begs that point. For the real question is whether there actually could be such a suspension or conflict. Hence to name its agent or source is not so much to fight on fair terms as to agree not to fight at all.

It does not follow that this qualification is unimportant. But its function is not what one might first suppose. Its real point is to give plausibility to events which, unlike the merely rare or surprising, would not otherwise be credible. It is to render possible events made impossible by a premature and perhaps parsimonious view of nature. In the case of miracles, God's role is like that of a gigantic wind introduced to account for waves on a normally glass-smooth body of water: it is to give the machine such a bash that it skips a cog; it is to overpower it so that it no longer behaves like a machine. And it is from this alleged divine interference with the course of nature that miracles as traditionally conceived have derived their religious, or better, their pseudo-religious force and significance.

This point can be put more positively. *Miracle*, in the properly supernatural sense, involves a cutting across what is or is thought to be the ways of nature or the real grain of the world. This is the core of the notion and, I believe, the root of the difficulty. Our point is that, so conceived, *miracle* cannot correctly and consistently describe any event, whether real or alleged. Or, to put the matter another way, that the attempt so to describe such events must inevitably lead to cramps and contradictions.

The third observation is more fundamental to our argument. This point concerns both *miracle* and *paradox* and, indeed, the notion of discrepant fact which is their common core. However, it can be stated most simply with specific reference to the first of these conceptions.

It follows from the traditional definitions of *miracle* that the question whether any event may be so described is really a question of the range or application of the concept of natural law. And one is tempted to think that this is simply a factual question. In fact, in one important sense, it is an historical one. More precisely, it is a question of how this latter concept is and, perhaps, must be understood.

There is no doubt that, at the beginning of the modern age, scientists saw natural law as limited in its range or application. Though they dared to see many of the ordinary and familiar events of the physical universe as due to natural causes they nevertheless assumed that at least certain events in the world were in fact due to the direct action of God. Thus they could, with perfectly good conscience, see these events as violations of law or as miracles in the supernatural sense. But now the situation is radically changed. Despite their doubtlessly justifiable reservations concerning the concept of natural law, the majority of contemporary scientists are convinced that this concept, if it is to be allowed at all, is and must be universal in its application. They assume that all events are lawlike (whatever that really means) or, at least, that they must be treated as such. They assume that every event can be shown to be an instance of some generalization, whether simple or statistical. This is why the scientist holds that there are no suspensions of natural law.

The nature and importance of this shift can be seen more clearly if we turn for a moment from the concept of natural law to a set of related categories. In Newton's age, for example, the world was conveniently divided into two mutually exclusive types of events. While some were accepted as supernatural and hence unintelligible, the majority of events were coming to be regarded as purely natural and hence as, in principle at least, open to human understanding. But the contemporary scientific view is much more radical. On essentially methodological grounds it discards the concept of the supernatural and resolves to treat all events as wholly natural. It proposes to treat all events as, in the final analysis, expressions of some one single pattern. Indeed, it regards any theory or hypothesis as adequate just to the extent to which it finds room for every conceivably relevant event.

A moment ago I suggested that the argument of this paper turns about the current scientific use of *natural law*. In fact, as has been suggested, it turns about a very far-reaching shift of which this is only the most notable expression. At the end of this paper I shall attempt to suggest very briefly that the extended use of terms such as natural law is justified and even necessary. For the moment, however, I content myself with two brief comments. First, unlike certain philosophers, every single scientist with whom I have discussed this question uses and reports himself as using such concepts in this sense. Second, I simply do not see how it is possible to do science or, better, how it is possible to "justify" it, on any other basis.

II

There are two main supernaturalist senses of *miracle*. They are: (1) an event involving the suspension of natural law, and (2) an event conflicting with our understanding of nature. We begin with the first of these senses.

The core of our objection is quite simple: the idea of a suspension of natural law is self-contradictory. This follows from the meaning of the term. Natural law is not, as has been widely supposed, a kind of code for nature having legislative and, perhaps particularly, prohibitive force. This is an outdated, untenable, and completely unscientific view. Natural laws bear no similarities to civil codes and they do not in any way constrain the course of nature. They exert no opposition or resistance to anything, not even to the odd or exceptional. They are simply highly generalized shorthand descriptions of how things do in fact happen. (This is misleading but will do for the moment). Hence there can be no suspensions of natural law rightly understood. Or, as here defined, *miracle* contains a contradiction in terms.

Once we understand natural law in this proper sense we see that such law, as distinct from our conception of it, is inherently inviolable. Hence anything which happens, even an apparent miracle, happens according to law. Or, negatively, no actual event could possibly violate a law of nature. Hence there can be no miracles in that sense of the term with which we are now concerned. Or, less misleadingly, it is in the nature of the case impossible that there should be an event which could be properly described by this term.

This contradiction may stand out more clearly if for *natural law* we substitute the expression *the actual course of events*. *Miracle* would then be defined as "an event involving the suspension of the actual course of events." And someone who insisted upon describing an event as a miracle would be in the rather odd position of claiming that its occurrence was contrary to the actual course of events.

But miracle (1) is not merely a contradiction in terms. It is also what might be called a contradiction in use. We see this in the plight of one confronted with an event conceived as a miracle. Such a person is faced with a dilemma. He can affirm the reality of this event and repudiate the "laws" which it violates. Or he can affirm these "laws" and repudiate the event. What he cannot do is to affirm both the event and the "laws" of which it is a violation. And this is to say that he cannot affirm that the event is both real and miraculous.

This can be put in a slightly different way. Belief in an event is belief that it happened. Belief that an event is a miracle is belief that it violates natural law. But, as already seen, such violations are impossible. Hence in respect of any alleged event we must choose between believing that it was actual and that it was not even possible. What we cannot believe, what it makes no sense to believe, is that any real or alleged event is a miracle in this sense. This, of course, is not to legislate for the course of events in the empirical world: it is simply to draw out the consequences of the self-contradictory nature of this concept.

So much for sense (1). What now of the more sophisticated (2) ("an event conflicting with our understanding of nature")? The clash posited by this sense is not within nature, as in (1), but rather between an event and our conception of nature. The idea of such a clash is not contradictory. Nor are such clashes either infrequent or always illusory. Hence there might be occasions on which *miracle* could be legitimately used in this sense. For example, I might use "x is a miracle (2)" to mean "x baffles me" or "I know no laws which could account for x." And if this was all I wanted to say, if my only concern was to register my lack of comprehension, then I would be as invulnerable as when I describe my dog as brown. But there are two relevant points. The original statement, so intended, deals not with the event as such but rather with my reaction to it; here the term *miracle* is being used expressively rather than descriptively. Secondly, in fact people do not use this term merely in this limited way. They do not use it simply to say that they do not know the appropriate law: they use it also to express their belief that there is no such law. They use it not merely to indicate their lack of comprehension but at the same time to insinuate their belief that the event cannot, perhaps should not, be comprehended. And if they did not so use it, it would be difficult to see how this term could have any distinctive meaning, much less the peculiar "religious" force it obviously has.

Though innocent when used expressively, this sense is contradictory when used to describe an event. One who so uses it implicitly affirms a conception of nature with which this is at variance. But he cannot believe both that the event happened and that the conception of nature with which it conflicts is adequate. In attempting to do so he necessarily contradicts himself. He is like the man who says, "Yes, this cat is white" then blandly adds, ". . . but I still hold that all cats are black."

Such a person may reasonably be asked to surrender either the historicity of the event or the conception of nature with which it conflicts. But as the proponent of an allegedly historical event he cannot well do the former. Nor, if the evidence was ever adequate, should he do so. He must therefore admit the inadequacy of the conception of nature with which he has been working. To do this is to put brackets around it, to withdraw it from circulation; it is to void or cancel the check. But in so doing he has cut the thread by which the whole structure was suspended and has repudiated the ground upon which he originally urged this description. The moral is clear: to affirm the historicity of an event is to destroy the ground for regarding it as a miracle.

Of course there is one other move open to our victim. Instead of surrendering his conception of nature as inadequate he might instead proclaim it a true and faithful copy. In so doing he locates the clash once again within the sphere of nature. But this brings him back to sense (1) and the impossible contradictions with which it is afflicted.

It appears then that one who insists upon describing an event as a miracle is faced with two equally impossible alternatives. If he attempts to represent the event as an instance of (2) he is forced by the logic of his position to withdraw the term completely. If he attempts instead to represent it as a case of (1) he forces himself to say that the event in question both is and is not a part of the actual course of events. Such a person is like a fly in a spider's web, he has ways around but no way out.

The conclusion of this matter can be stated "paradoxically": all the properly descriptive senses of *miracle* are logically improper. And, though this is to anticipate, when we feel the need to describe an event as a miracle we should construe this to mean that we have already misinterpreted the nature of reality.

III

At first sight *paradox* seems quite unlike *miracle*. That is perhaps because this term is ordinarily applied to statements and only derivatively to events. But the difference is more apparent than real. "Paradoxical" statements, at least as here considered, are simply affirmations of "paradoxical" (or, in a wide sense, "miraculous") events. This permits us to argue in the same vein as before and even to make some use of the spade-work already done. It also allows us to concentrate upon demonstrating that there can be no actually paradoxical events; for if this is so, neither can there be paradoxical statements. Or, more correctly, neither can there be statements which are both true and paradoxical.

The connection between paradox as statement and as event is evident in the main senses of this term. Arranged so as to bring out their correspondence with the various senses of *miracle* they are: (1) a person or thing (event) conflicting with true or adequate notions of what is reasonable or possible; (2) an event conflicting with preconceived notions of what is reasonable or possible; (3) a statement actually self-contradictory or absurd; and (4) a statement seemingly self-contradictory or absurd. We shall begin with (1), proceed to (2) and, finally, refer briefly to their statement counterparts.

Paradox (1) postulates a clash between an event and one or more true or adequate background conceptions. In so doing it assumes a break or rupture within the fabric of being. Hence it is virtually on all fours with miracle (1). And it too is a contradiction in terms involving those who use it to describe events in logical confusions and perplexities.

All this is apparent from the definition. "A true notion of the possible," to abbreviate the *definiens*, is just a notion which has room for all events. To speak of "an event conflicting with true notions of the possible" is contradictory in the same way and to the same extent as speaking of "an event involving the suspension of natural law." Both are inherently self-contradictory.

By the same token, it is impossible that an event should conflict with true notions of the possible. Conflicts between "events" and "true notions of the possible" can occur only if the event is a mere fiction or the allegedly true notion is in fact false. Between real events and true notions there can be no conflict. Hence it is impossible that there should ever be a paradox in this sense.

As with the corresponding sense of miracle, one becomes involved in contradiction when he attempts to describe an allegedly historical event as a paradox in this sense. This can be shown quite easily in terms of the original definition. One who maintains the reality of an event implicitly rejects as inadequate those notions with which it conflicts. But in describing it as a paradox he also affirms these same notions. In short, in so describing an event he necessarily contradicts himself.

Paradox (2) (an event conflicting with preconceived notions of what is reasonable or possible), like the corresponding sense of miracle, seems much more cautious. The clash which it postulates is not within nature but rather between an event and certain of our preconceptions. Thus, unlike (1), it leaves the realm of being whole and intact. Hence it is not on the face of it absurd.

Now it is of course true that things can and do happen which conflict with our preconceived notions and ideas. In this sense things happen which we might want to speak of as paradoxes in this sense. So understood, the term may be given an expressive use. But the person who attempts so to describe an allegedly real event is already in serious logical difficulties. Such a person is faced with a dilemma. He can affirm the reality of this event and repudiate as inadequate the preconceptions with which it conflicts. Or he can affirm these preconceptions and repudiate the event. What he cannot do is to affirm both the event and the preconceptions with which it conflicts. And this is to say that he cannot believe that an event is both real and paradoxical.

Again. In alleging the event he repudiates the conflicting conception. In describing it as a paradox he affirms that conception. But he cannot both repudiate and affirm this conception. Therefore he must sacrifice either its historical or its paradoxical character.

To allege the occurrence of something conflicting with one's preconceived notions is to compel oneself to revise those notions. And this is to destroy the ground for regarding the event as a paradox. So to describe an event is a move we can make only once in a given game, and it is a move which necessarily defeats itself.

The attempt to represent an event as a paradox involves the speaker in an impossible contradiction. If he treats it as an instance of (2) he tacitly accepts as adequate certain conceptions which admittedly clash with an allegedly real event. If, instead, he treats it as an instance of (1) he is saying both that it happened and that it was not even possible; both that it is part of the real world and that, in some odd way, it is not part of that world. Whichever sense he adopts he necessarily contradicts himself. This is to say that he cannot so describe any real event. Or, what comes to the same thing, that there are no events which are both real and inherently paradoxical.

Senses (3) and (4), as already stated, are respectively the statement counterparts of those senses just considered. It follows then that no affirmation of an event can be, or be consistently treated as, an instance of either of these senses. Or rather, since in the case of events the alternative is between being real and paradoxical, it follows that no affirmation of an event can be both true and paradoxical. Hence, if a statement is actually paradoxical it must also be false.

IV

It is now clear that each of the various senses of *miracle* and *paradox* is contradictory and without counterpart in the real world. The next step in the argument is to show that this is also true of a conception which is common and fundamental to both these terms. The conception in question is that of an irregular or discrepant fact.

Now no one will deny that events can be discrepant with pictures and conceptions of the universe which have thus far seemed entirely adequate. But they cannot be discrepant with a true or adequate conception of the universe nor indeed with the actual course of the universe itself. The reason is quite simple. By definition, the actual course of events includes what we might call the actual course of every event. Events may be and often are strange and disturbing. But they are never "outside nature." They are never "out of line" with the real way of the world. They are never discrepant with the whole. This is a matter of definition and no intervention, whether divine or diabolical, can alter it.

By the same token no one can bring himself to believe that a real event is actually discrepant with an adequate conception of the universe. And the reason is simply that *adequate conception* cannot mean anything less than one which takes account of, embraces, and comprehends all real events. And this is equally true in the case of alleged events. One cannot maintain both the reality of an event and the adequacy of the conception of nature with which it is in conflict.

Events come to be treated as discrepant for a variety of reasons the most important of which are humanity's impatience, lust for simplicity, and relative poverty of experience. The role of each is apparent in the following example. Imagine a universe in which there are but x events. Of these, x-1 are instances of persons drowning upon prolonged submersion in water. The remaining event is one in which an Indian holy man is still alive after eight days in this state. The proponents of miracles and paradoxes want to say that this single event happened contrary to the way in which things happen, that it is discrepant with the real course of events. But this is at once arbitrary and groundless. Both drowning and non-drowning are part of the way things happen in this universe. The holy man's survival is admittedly strange and surprising, and it will conflict with current conceptions of the possible. But the fact that it happened can only mean that these conceptions are premature and inadequate. It may well be an exception to a rule but it is not, any more than are all the remaining events, a violation of natural law. And that law being what it is, it is impossible that it should be.

Another example. One can say, "He is taller than any other member of his family." But he would be guilty of a confusion if he said, "He is taller than any member of his family." Yet when we accept an event as discrepant this is how we talk.

Again, one can imagine an adoring wife saying to a wayward husband, "Oh dear, John, I wish you would be more like yourself." But when people talk this way it is difficult to avoid feeling that they ought instead to revise their theories.

Now, to bring these claims together. There are no events which are discrepant with, or outside, the course of nature; this is because all events are part of nature and because the manner of their occurrence is necessarily a part of nature's actual course. There are no miracles or violations of natural law: this is because *natural law* is simply an expression for the way in which things actually happen. There are no paradoxes or events in which things act contrary to their real nature: this is because the essence or nature of a thing is simply a shorthand for the various ways in which it can act. Finally, a related claim for which the ground has not yet been turned; there are no irrational events because, in the last analysis, *rational* can only mean something like *the way in which things really are.*

It is alleged that there are neither discrepant events, miracles, nor paradoxes. But this is not to deny the historicity of those events which have been traditionally so described. That is an entirely different question which can be settled only upon the basis of the relevant evidence. The question whether there are miracles and paradoxes is not a question of whether certain events happen. It is, rather, a question of how we are to describe events which do happen. *Miracle* and *paradox* are not names of events (as "the raising of Lazarus" is the name of an event), but mythological categorizations which we impose upon events. In such cases the question is whether the framework makes any sense. We have rejected this one on the ground that it is marred by logical flaws and contradictions.

V

The argument is finished but it would be sanguine to suppose that the supernaturalist has been convinced. His belief in the possibility of discrepant events is grounded in his belief in a special divine agency, and because no attention has been paid to such an agency, he is almost bound to feel that the whole question has been begged from the outset. The final step then is to show that this omission has not prejudiced the matter. It is clear that this really requires at least a description and criticism of the supernaturalist world view and its relation to the rise and presuppositions of science together with some account of the genesis, nature, and extent of our knowledge of the external world. But space permits no such excursion and we must settle for rather blunter treatment.

The first point is that it makes no sense to speak of events as being *specially caused* by God. By this is meant that there is no meaning which can be attached to such a conception. At least, granted the supernaturalist account, there is no such meaning. If John Jones is a painter of a picture it makes no sense to say that he is the special painter of a particular part of it. And if, as the supernaturalist professes, God is the cause of all events, it makes no sense to say that he is the special cause of one or more particular events. This suggests that the supernaturalist speaks of special cause in this connection because he does not really regard God as in any sense the cause of regular events. Certainly it is only on such an assumption that it is possible to attach any meaning to the phrase *special cause*. But what is then intended ought rather to be described simply as a cause.

Secondly, even if it were possible to attach some specific meaning to this phrase, the situation would not thereby be altered. For there is in fact no way of telling whether an event has been specially caused or not. To see this, it is necessary only to consider an earlier example. Was the Indian holy man's survival a specially caused event or was it merely rare? There is no way of answering this question (and perhaps there should be no way of asking it). The Christian may want to say that he knows that the miracles of the Christian faith were specially caused by God. But how does he know? And is it in some way different from that in which the disciple of the holy man knows that his master's feat was specially caused by God?

In this connection it may be relevant to note that the events described as specially caused are invariably of extremely rare occurrence. The traditional explanation is that divine intervention is and ought to be rare. But there is another sort of explanation which is rather more enlightening. People so account for such events because their conception of the world is simply a transcript of the recurrent and familiar, because in the light of such a transcript these events seem properly mysterious, and because they are sure that the ways of the Lord must be hidden and mysterious. Anyone who doubts this has only to ask two questions. Which is more strange and obscure in its mode of operation: that Christ should have turned water into wine, or that Beethoven should have eaten bread and produced symphonies? And would the latter ever be described as a miracle?

This brings us to the final point. Even if certain events were specially caused and identifiable as such, they would not on that account be discrepant. What appears to be a breaking of divine habits must really be a part of those habits. What seems to be an instance of special divine activity must simply be a case of divine activity. Rare events are important not because they are discrepant, but because they are and must be treated as coherent. This is the lesson which the supernaturalist has to learn from his stepchild, the scientist.

VI

Our argument has assumed the contemporary extended use of *natural law* and, implicitly at least, that of certain related expressions. It now remains to suggest, however briefly, that this use is justified and, indeed, necessary.

Obviously the resolve to treat the world as a single whole would be justified if we knew that, in some quite simple and straightforward way, the world was actually such. For example, it would be justified if we knew that it was the expression of a single though complex thought in the mind of God or, again, that it was the result of instantaneous creation. Or at least we can imagine that it might be thus justified. But, fortunately or not, we seem to lack any such justification. We do, however, have some important support in the "shape" of the human mind or, rather, what appears to be the conditions of its operation. To put the matter in the briefest possible way, whether we think of the actual conduct of our scientific investigations or the criteria in the light of which we assess their results, we feel bound to reject as premature and inadequate any hypothesis or explanation which does not account and make room for every one of the events with which we are concerned. This is not to say that such coverage is identical with truth and certainly it is not to suggest that we could somehow know that there is such a connection. It is simply to make the point that the mind can proceed with its investigations and with the assessment of its own results only to the extent that it does in fact treat the world as a single and coherent whole. In short, seeing the world in this way is a condition without which the mind could not do its work. In this sense then this resolve is justified and even necessary.

It is in this light that we must see the scientist's resolve to treat all events as subject to natural law. This does not mean that he insists that events should conform to some conception he already has. Nor does it mean that he disregards those which he has not yet been able to fit within such a conception. Rather, it means that he has resolved to view all events in this light. For him, *law* is a slogan; it is the way in which he proposes to look at the world. His acceptance of all events as expressions of natural law is the way in which he guides himself in his attempt to discover the real content of this conception. It is therefore essential that he refuse to treat any event as discrepant. This is not to say that certain scientists have not so treated events upon convenient occasion. It is only to say that when they have done so they have ceased to be scientists. The mark of science is not some theory concerning the adequacy of efficient causes but the resolve to see every event as coherent with the whole.

Now the true religious believer is also interested in the events of the world. He sees these as instances of divine activity and he knows that it is only by carefully examining them that he can even begin to understand either the world or the divine activity of which, in his view, it is an expression. The significance of the preceding argument is obvious. In the determination of these conceptions nothing will serve him better than the resolve to treat every single event as coherent and revelatory, and nothing will mislead him more than to conceive some events as miracles and paradoxes.

McGill University

Received September 22, 1966

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 4, Number 4, October 1967

VII. MOORE'S NATURALISTIC FALLACY

DAVID P. GAUTHIER

> 'Good,' then, denotes one unique simple object of thought among innumerable others; but this object has very commonly been identified with some other—a fallacy which may be called 'the naturalistic fallacy.'[1]

G. E. Moore's alleged discovery of the naturalistic fallacy has decisively altered moral philosophy in the English-speaking world. So much is uncontroversial, but even to ask what is the fallacy invites disagreement.

I. *What Does Moore Refer to by the Phrase "Naturalistic Fallacy"?*

An examination of Moore's explicit statements[2] suggests the following characterization:

(*a*) the word "good" denotes an object—in particular, an object of thought;

(*b*) this object is simple, i.e., not complex, not capable of being analyzed into parts;

(*c*) this object is unique, i.e., not identical with the object denoted by any other word;

(*d*) this object is, however, often identified with the complex object denoted by some expression, or with the simple or complex object denoted by some other word;

(*e*) this identification is expressed by the claim that the word "good" has the same meaning as the other word or expression; i.e., that the other word or expression provides a definition of "good";

(*f*) this identification is the naturalistic fallacy.

The naturalistic fallacy is often considered to be the fallacy of improper definition—of defining a word which cannot be defined. Moore himself gives license to this view in insisting that "good" is indefinable. But his account makes clear that the fallacy is more properly considered to be the fallacy of improper identification—of identifying two objects which are distinct.[3]

In sect. 13 of ch. I of *Principia Ethica* Moore undertakes to demonstrate the naturalistic fallacy. Before examining the content of his argument, we may show that its form confirms our account of the fallacy.

Moore never considers that he might be wrong in supposing that "good" denotes an object—he never questions (*a*). In sect. 13 he considers only (*b*) and (*c*). If (*b*) is false, then "good" denotes a complex object, and so is definable. By "definition" Moore means "the analysis of an object of thought" (p. xiii); "a definition states what are the parts which invariably compose a certain whole" (p. 9). Only complex objects, and all complex objects, are definable; in defending (*b*), then, Moore defends his claim that "good" denotes what is simple and so indefinable.

If (*c*) is false, then "good" denotes an object denoted by some other term. Moore expresses this possibility in a manner which critics have understandably found misleading; he says that *good* "means nothing at all, and there is no such subject as Ethics" (p. 15). But his discussion makes amply clear that he means that "good" has no distinct meaning, that "good" does not stand for a unique object or property. Further, in saying that there is no such object as Ethics, Moore means that there is no distinctive subject; if, for example, the object denoted by "good" is the object denoted by "pleasure," then Ethics is a part of psychology.

In sect. 13, then, Moore defends (*b*) and (*c*). If (*a*), (*b*), and (*c*) be accepted, it follows that the identification mentioned in (*d*) must be mistaken, and so may correctly be considered a fallacy.

---

[1] G. E. Moore, *Principia Ethica* (Cambridge, 1903; reprinted 1929), p. xiii. Otherwise unidentified quotations are from this work, with the page number(s) in parentheses.

[2] For example, *ibid.*, pp. 58, 73.

[3] On this point cf. H. J. Paton, "The Alleged Independence of Goodness" in P. A. Schilpp (ed.), *The Philosophy of G. E. Moore* (New York, 1952), p. 133. Also cf. Abraham Edel, "The Logical Structure of G. E. Moore's Ethical Theory," *ibid.*, pp. 146–147.

II. *How Might One Show That What Moore Refers to by the Phrase "Naturalistic Fallacy" Is a Fallacy?*

Let (*b*) be denied. Then, Moore says, "whatever definition be offered, it may be always asked, with significance, of the complex so defined, whether it is itself good" (p. 15). Rather than consider Moore's rather difficult example, we shall introduce a purely abstract definition:

"good" =df '*x*'

Moore's position[4] is that:

(i) the question "Is it good that *A* is *x*?" is significant;

(ii) the question "Is it good that *A* is *x*?" is not equivalent to "Is it *x* that *A* is *x*?";

(iii) "That *A* should be *x* is good" is not equivalent to "That *A* should be good is good."

From (ii) and (iii) it is supposed to follow that the proposed definition fails.

What force does this argument have? To answer this question, we must have a clear understanding of the expression "It is good that *A* is good." But neither I, nor to my knowledge anyone else, has such an understanding. Indeed, if some *x* seemed to me to provide a plausible definition of "good," then I should be willing to interpret "It is good that *A* is good" in such a way that the equivalences rejected by Moore would in fact hold. And this would be to set Moore's argument on its head.

At this point, however, it might be objected that I have taken Moore's argument at face value, assuming that he meant what he said. Perhaps one should assume that Moore, confused by his own example of a proposed definition of "good," actually meant to argue the point that, given any definition of "good," say '*x*' then it is possible significantly to ask "Are things which are *x* good?," and not be asking about the meaning of the words. I shall examine this argument later.

In discussing the denial of (*c*), Moore supposes that someone equates good and pleasure, so that "the proposition 'Pleasure is the good' does not assert a connection between two different notions, but involves only one, that of pleasure, which is easily recognized as a distinct entity. But whoever will attentively consider with himself what is actually before his mind when he asks the question 'Is pleasure . . . after all good?' can easily satisfy himself that he is not wondering whether pleasure is pleasant" (p. 16).

But this argument is of no force at all. For suppose someone offers the definition:

"brother" =df "male sibling"

Certainly we understand what is meant by doubting whether a brother is always a male sibling, although if we are satisfied with the definition we shall no longer doubt this. And certainly if we ask "Is a male sibling after all a brother?" we are not wondering whether a male sibling is a male sibling. What we are doubting, or wondering about, is whether the object denoted by the word "brother" is the same as the object denoted by the expression "male sibling."

From Moore's own arguments, I turn to the test of self-contradiction.[5] Given the definition:

"good" =df '*x*'

it follows that the statement "*A* is good but not *x*," and the statement "*A* is *x* but not good" are self-contradictory. But, it is urged, for any *x*, where *x* is a term denoting some natural property simple or complex, these statements are not self-contradictory, even if they are always false. Hence the definition fails.

It is clear that this argument depends on the assumption that the definition of "good" provides a verbal synonym for "good." But for Moore, a definition does not express verbal synonymy, but rather provides an analysis into parts. Moore does not give a full example of a definition, but he does provide a partial definition of horse (not "horse"). "When we say . . . 'The definition of horse is "A hoofed quadruped of the genus Equus," ' we may, in fact, mean three different things." The first two are the arbitrary verbal definition and the verbal definition proper; Moore insists that such definitions may be given for "good," but that he is not concerned with them. In the third, important sense, "We may mean that a certain object, which we all of us know, is composed in a certain manner: that it has four legs, a head, a heart, a liver, etc., etc., all of them arranged in definite relations to one another" (p. 8).

A real definition—for that is what Moore has in mind—is not intended to state a verbal equivalence. Thus even if every definition of "good" fails the self-contradiction test, it does not follow that good is indefinable or simple.

---

[4] Moore, *op. cit.*, pp. 15–16.

[5] Cf. Paul W. Taylor, *Normative Discourse* (Englewood Cliffs, 1961), p. 243; and Bernard Mayo, *Ethics and the Moral Life* (London, 1958), p. 71. The open-question test, which Moore may have been trying to formulate in his discussion of (b), reduces to this.

Does the self-contradiction test show that the object denoted by "good" is unique? It would seem not. It is true that the substance denoted by "salt" (in its ordinary use) and the substance denoted by "sodium chloride" are the same, but "salt" and "sodium chloride" are not synonyms. We discover that salt is sodium chloride by chemical analysis. For chemical purposes it may be useful to define salt as sodium chloride, but that is a real, not a verbal, definition, which does not deny that it is a contingent fact that this stuff in the salt-cellar is a chemical compound whose molecules contain one atom of sodium and one of chlorine.

Similarly, then, if the naturalist makes the claim that the object denoted by "good" and the object denoted by "pleasant" are the same, it will not be possible to show him to be wrong, just by showing that "good" and "pleasant" are not synonymous. That two denoting-expressions denote the same thing shows nothing about the sense or meaning of the expressions.

At this point we must consider an important difficulty. It is true, of course, that two expressions may refer to the same object or *substance* without being synonymous. But Moore supposes that "good" denotes a particular object of thought, a *property*. Is it possible that another term might denote the same property, yet possess a different connotation, and so not be synonymous with "good"? Consider the following account of the naturalistic fallacy:

> The naturalistic fallacy is the assumption that because the words 'good' and, say, 'pleasant' necessarily describe the same objects, they must attribute the same quality to them. We might, with Mill, call the objects to which a term is applicable the denotation of the term, and the characteristics which an object must have for the term to be applicable to it, the connotation of the term. What the man who commits the naturalistic fallacy fails to realize is that 'good' and some other adjective may denote or be applicable to the same things, and yet not connote the same quality, i.e., describe the same things in the same way.[6]

Prior's use of "denote" is quite different from Moore's (a point Prior does not mention). In Prior's terminology, the naturalistic fallacy is committed if the property connoted by the word "good" is confused with the property connoted by some other word or phrase. And this confusion is expressed by the claim that the word "good" has the same meaning as the other word or expression.

Hence the naturalist claims that the property connoted by "good" is the property connoted by '*x*.' And it is possible to show him to be wrong if it is possible to show that, for every *x* connoting a natural property, "*A* is good but not *x*," and "*A* is *x* but not good" are not self-contradictory. The self-contradiction test, if it works, shows that what Moore refers to as the naturalistic fallacy is, under Prior's interpretation, a fallacy.

Prior's account of the naturalistic fallacy requires the supposition that "good" and "pleasant" refer to the same property if and only if "good" and "pleasant" mean the same. On this view, we may note, the question of the identity of mental and physical characteristics cannot seriously arise. For since it is clear that "pain," for example, does not *mean* "stimulation of nerve-fibres," it follows that pain cannot *be* the stimulation of certain nerve-fibres.

This example should suggest my dissatisfaction with Prior's supposition. It permits the solution of too many problems too easily. But I want to consider another example, both because Moore considers it, and because it sheds considerable light on the importance of the naturalistic fallacy.

On Prior's supposition the property referred to by "yellow" cannot be identical with the property referred to by some phrase concerning the reflection of light-vibrations of a certain wave-length, since these expressions are by no means synonymous. Yet do we want to treat the yellowness of an object as a property different from its capacity to reflect certain light-vibrations? I should not want to; if I were to ask someone to state two properties of sulphur, I should not accept the answer "It is yellow and it reflects light of such-and-such wave-length." But I do not know how I should persuade someone who disagreed with me on this point, and I am not at all sure that the difference between us would not be merely verbal. He speaks of two properties; I speak of one; but we agree on all of the facts. In particular, we agree that the relation between being yellow and reflecting light of the appropriate wave-length is logically contingent but physically necessary.

Among those who do disagree with me we must include Moore. Moore suggests that someone might try to define yellow "by describing its physical equivalent," i.e., the light-vibrations. But he insists, correctly, that "those light-vibrations are not themselves what we mean by yellow." And from this he concludes, "The most we can be entitled to say of those vibrations is that they are what corresponds in space to the yellow which we actually perceive" (p. 10).

---

[6] A. N. Prior, *Logic and the Basis of Ethics* (Oxford, 1949), pp. 1–2.

Thus although Moore and Prior express themselves differently they would seem to agree that two property-words refer to the same property if and only if they mean the same. If so, then the alleged naturalistic fallacy is a fallacy if and only if, for every term *x*, either *x*=good or "*A* is good but not *x*" is not a self-contradiction.

Note that in this form it is highly unlikely that the test of self-contradiction should prove successful. Some restriction must be placed on the range of *x*. The obvious one—although not the one which Moore accepts—is to insist that *x* must refer to an observable property.

III. *What Follows, of Importance to Moral Philosophy, from the Claim that What Moore Refers to by the Phrase "Naturalistic Fallacy" Is a Fallacy?*

Let us suppose that Moore is right in supposing good to be a property. Now let us suppose that someone else claims to have discovered the psychological equivalent of this property—claims, for example, to have discovered a correlation between a thing's being good and its being the cause of pleasant feelings.

Has the naturalistic fallacy been committed? No. The property referred to by "good" has not been identified with the characteristic referred to by "cause of pleasant feelings." The self-contradiction test is clearly of no relevance to whether the alleged discovery is correct.

But now suppose that someone claims, on the basis of this discovery, that the property referred to by "good" is the same as the property referred to by "cause of pleasant feelings." Has the naturalistic fallacy been committed? Yes. For if it is agreed that "good" does not mean "cause of pleasant feelings," then by supposition "good" and "cause of pleasant feelings" cannot refer to the same property.

The triviality of the naturalistic fallacy should now be apparent. For the difference between taking "cause of pleasant feelings" to refer to the psychological equivalent of the property referred to by "good," and taking "cause of pleasant feelings" to refer to the property referred to by "good," is purely verbal. If we choose to say, as Moore seems to and as Prior does, that two terms refer to the same property if and only if they mean the same, then we shall contradict ourselves if we identify good with the cause of pleasant feelings, but the choice, and so the contradiction, are of no consequence.

The claim that what Moore refers to as the naturalistic fallacy is a fallacy thus in no way impedes the naturalist in developing his ethical theory. The naturalist claims that the property referred to by "good" has an observable equivalent —a physical or psychological equivalent. Using this equivalent he may then develop a naturalistic, scientific theory of good and bad. His procedure may resemble that of the physicist, who shows that light-vibrations are the physical equivalent of color, and then proceeds from that equivalence to develop a theory of color. Or his procedure may resemble that of the physiologist, who claims that certain stimulations of nerve-fibres are the physical equivalent of pain, and proceeds then to develop a physiological theory of pain.

The physicist may speak loosely of light-vibrations as defining color. The physiologist may speak loosely of stimulations of nerve-fibres as defining pain. These accounts may reveal a confused concept of definition, but such confusion provides no ground for rejecting the underlying theories of the physicist and physiologist, although it may provide ground for rephrasing the formulation of these theories. Similarly the naturalistic fallacy, if it is a fallacy, is at best a ground for rephrasing the accounts of some naturalists, not for rejecting the theories they propose.

The possibility of confusing statements of physical or psychological equivalents with statements of verbal equivalents is of some importance. If a naturalistic moral theorist makes this confusion, he may argue: "Pleasure is the natural equivalent of good. So 'good' just means 'pleasure'. So anyone who objects to my theory is guilty of verbal confusion; he is just misusing the word 'good' (or 'pleasure')." This obviously bad argument may lead the naturalist to reject or ignore objections to his taking pleasure to be the natural equivalent of good, by assuming the verbal equivalence of "good" and "pleasure" and then showing that the objection contains a contradiction.

However, an anti-naturalistic moral theorist may make the same confusion, and argue equally badly: "If pleasure were the natural equivalent of good, then 'good' would just mean 'pleasure'.

But 'good' does not mean 'pleasure'. Therefore pleasure is not the natural equivalent of good—anyone who supposes it is must be committing the naturalistic fallacy." Both arguments depend on the utterly false assumption that if *x* is the natural equivalent of *y*, '*y*' means '*x*'. Moore was well aware that this is false, as his discussion of the physical equivalent of yellow shows.

Moore's naturalistic fallacy is trivial. But it is often supposed that the fallacy consists, not in identifying the property referred to by "good" with some natural property, but rather in alleging that some natural property is the natural equivalent of the property referred to by "good." And this is by no means trivial, for it is incompatible with any naturalistic theory of ethics. To distinguish this view I shall refer to it as the (alleged) *naturalistic error*. The naturalistic error is to suppose that good has a natural equivalent.

Of course I am not claiming that this is an error. The label is intended only to create some parallel with the naturalistic fallacy. More neutrally, it might be termed the *naturalistic assumption.*

Before turning to the naturalistic error, let me point out that there are other reinterpretations of the naturalistic fallacy which need not be trivial. For example, it is sometimes supposed that the fallacy lies in denying the semantic autonomy of ethical and evaluative terms, in supposing that they are eliminable from discourse without significant loss. Or it is supposed that the fallacy lies in supposing that ethical and evaluative conclusions are logically entailed by statements of natural fact. But neither of these will concern me here.

IV. *How Might One Show that What I Am Referring to as the Naturalistic Error Is an Error?*

None of the arguments advanced by Moore, nor any test which rests on verbal equivalence such as the test of self-contradiction, will show that the naturalistic error is an error. On the contrary, Moore did not *consistently* suppose the naturalistic error to be an error; for he insists that it is not an error to define the good—that which is good. But if we can define the good, then "some true proposition of the form 'Intelligence is good and intelligence alone is good' can be found; if none could be found, our definition of *the* good would be impossible" (p. 9).

Now such a proposition would not provide a definition of the good if we were to suppose that the connection it asserted between intelligence and good were merely coincidental. But of course the connection is not logically necessary. I take it, then, that such a proposition would state that intelligence is the natural equivalent of good.

Moore's primary objection to naturalists, then, is not that they seek an equivalent for good, but that they confuse this task with seeking a definition of good. Or, that they confuse seeking a natural equivalent with seeking a verbal equivalent. Whether Moore is right about this seems to me very doubtful, but he is right at least to this extent. Many naturalists, and certainly some of those whom he criticizes, fail to see that such a claim as that pleasure is the equivalent of good requires to be established. The physicist does not just say that light-vibrations are the physical equivalents of colors; he shows this. The physiologist does not just say that the stimulation of certain nerve-fibres is the physical equivalent of pain; he seeks to show this. Similarly, the naturalistic moral theorist cannot just say that pleasure is the psychological equivalent of good; he must show this. And naturalists have often failed even to attempt to show this, or whatever other equivalence they suppose.

But if Moore is objecting primarily to the way in which naturalists do their job, is he objecting to naturalism itself? From what I have said, it would seem not; if Moore supposes it possible to find a natural equivalent of good, then he seems to be a naturalist himself.

This apparently paradoxical suggestion must be dismissed. Moore's search for a natural equivalent is fruitless. As he says,

> . . . if we set ourselves seriously to find out what things are good, we shall see reason to think . . . that they have *no* other property, both common and peculiar to them, beside their goodness—that, in fact, there is no criterion of goodness (pp. 137–138).

The consistency of this with "I do most firmly believe that some true proposition of the form 'Intelligence is good and intelligence alone is good' can be found" (p. 9) I leave to others to worry over.

Apart from this, a fundamental difference between Moore and the naturalists becomes evident if we ask the question which Moore accuses the naturalists of neglecting: how do we show that something is the natural equivalent of good?

A naturalist might insist that it must be answered by using empirical techniques which are in principle similar to those used to show light-vibrations to be the physical equivalent of color, or stimulation of nerve-fibres the physiological equivalent of pain. To apply these techniques, however, we must first know what things are good, just as we must first know what things are colored, or when we are in pain. But how do we know what things are good?

Moore insists that "no relevant evidence whatever can be adduced" (p. viii) to show what kinds of things are good in themselves. All propositions which state that certain things are good in themselves, and so give partial definitions of *the* good, are such that "from no other truth, except themselves alone, can it be inferred that they are either true or false" (p. viii). It is in this sense that Moore calls such propositions "Intuitions." "I mean *merely* to assert that they are incapable of proof; I imply nothing whatever as to the manner or origin of our cognition of them" (p. x).

The naturalist will not accept this account. But it surely parallels the account he does accept of how we know what things are colored, or how we know when we are in pain. For in these cases we no not appeal to evidence. We do not infer such propositions as "This fire-engine is red," or "I am in pain" from other truths. In Moore's sense they too are intuitions.

At one point our argument seemed to be turning Moore into a naturalist. Now it seems to be turning naturalists into non-naturalists. For if the naturalist is to maintain the parallel he holds exists between ethics and the other sciences, must he not agree with Moore that we have no evidence for claiming that certain things are good—that we just "see" good things?

Of course not. The naturalist is undoubtedly right to reject Moore's view of how we know certain things are good. But in rejecting it, he must also reject the parallel we have developed between finding a natural equivalent for good and finding a physical equivalent for color. And how is he then to determine what things are good? He cannot *assume* a natural equivalent for good and use it to determine what things are good; for as I have argued, he must first know what things are good, in order to determine the natural equivalent.

Hence we can see why Moore supposes the naturalist must confuse the task of finding a natural equivalent for good with that of defining good. For if the naturalist begins by saying that good is pleasure, and treats this neither as the statement of an intuition nor as the statement of a conclusion which requires demonstration, than what else can it be but the statement of a definition?

We may now answer the question: how might one show that the naturalistic error is an error? The naturalistic error, as I have characterized it, is the claim that good has a natural equivalent. Now one might argue that if good has a natural equivalent, then it must be possible to show this in a manner similar to showing that color has a physical equivalent. But this requires that we know, without evidence, either what things are good, or certain further truths which entail conclusions about what things are good. And this is absurd. Hence good has no natural equivalent.

To reject this conclusion, one must deny the conditional premiss that, if good has a natural equivalent, it must be possible to show this in a manner similar to showing that color has a physical equivalent. What alternative might be proposed?

Briefly, and dogmatically, the following. One shows that color has a natural—physical—equivalent by showing that there is a correlation between certain experiences (seeing colors) and certain physical states of affairs. But one shows that good has a natural equivalent—if it has one—by showing that there is a correlation between certain utterances (containing "good") and certain states of affairs. The naturalists must rest their case on semantic analysis.

University of Toronto

Received March 8, 1967